Gender domain adaptation for automatic speech recognition task

Sokolov Artem\textsuperscript{1}[0000-0002-3627-8683] and Savchenko Andrey\textsuperscript{1}[0000-0001-6196-0564]

HSE University, Nizhniy Novgorod, Russia
\{artsokol, avsavchenko\}@hse.ru

Abstract. This paper is focused on the finetuning of acoustic models for speaker adaptation based on a given gender. We pretrained the Transformer baseline model on Librispeech-960 and conduct experiments with finetuning on the gender-specific test subsets. Our approach leads to 5\% lower word error rate on the male subset if the layers in the encoder and decoder are not frozen, but the tuning is started from the last checkpoints. Moreover, we adapted our general model on the full L2 Arctic dataset of accented speech and fine-tuned it for particular speakers and male and female genders separately. The models trained on the gender subsets obtained 1-2\% higher accuracy when compared to the model tuned on the whole L2 Arctic dataset. Finally, we tested the concatenation of the pre-trained x-vector voice embeddings and embeddings from conventional encoder, but its gain in accuracy is not significant.

Keywords: automatic speech recognition, speaker adaptation, deep neural networks, gender-based acoustic models.

1 Introduction

State-of-the-art speaker-independent automatic speech recognition systems (ASR) show very accurate results when they are trained on large datasets with various speaker samples. However, the typical challenge even of very deep models is the dependance of recognition quality on acoustic characteristics of particular speaker. Indeed, it is practically impossible to collect the training dataset that contains enough hours of voice examples for all age ranges, genders, and accents for a certain language. The adaptation to a voice of particular speaker is possible, but usually requires much data.

The issue could be relaxed if a model is adapted to the group of speakers as it typically done for recognition of accents, namely, by using additional training of a model on a small dataset of accented data, that makes it possible to significantly increasing the recognition quality \cite{16}.

If we consider acoustic models based on deep neural network, the same approach but for gender groups is not studied enough. Indeed, gender is an important speaker variability factor. Females and males obviously have differences in their voices, Consonant noise and vowel formants of female speakers tend to be located at higher frequencies. Phonation type also seems to depend on the speaker’s gender. Female voices are often considered more breathy than male voices \cite{1}. Significant differences were...
present between genders in the distribution of energy throughout the analyzed frequency values, taken from voice samples [2].

These facts could be considered to reduce the error rates separately for male or female gender group, if a gender of a speaker is assumed to be known. Acoustic model could be adapted by fine-tuning on the voice samples from only one gender. As a result, ASR engine will contain several acoustic models. An appropriate model is chosen depending on the gender or age of the speaker. Nowadays, gender recognition and age prediction tasks are actively studied either for video [3, 4] or audio modalities [5, 6].

Thus, in this paper we examined the tuning of acoustic model for the gender domain. We used a neural network-based acoustic model and conduct experiments with two English datasets. Due to cross-gender voice differences are not equal for all languages [7] we choose datasets to cover native and not native speakers. The rest of the paper is organized as follows. Section 2 contains brief discussion of related literature. We discuss an approach to fine-tune acoustic models in Section 3. Section 4 contains experimental results for LibriSpeech and L2 Arctic datasets. Finally, concluding comments are given in Section 5.

2 Related work

Numerous studies have explored non-linguistic features differences between females and males and approaches to adapt various speech systems for groups based on gender, age, and accents [8, 9, 10]. Speaker adaptation techniques like maximum a-posteriori (MAP) and maximum likelihood linear regression (MLLR) are applied to ASR systems based on Gaussian mixture models (GMM) and Hidden Markov Models (HMM) [11, 12].

Acoustic models for gender domains could also be trained from scratch. The paper [13] reports accuracy improvement for using gender-based GMM for parliament speech recognition. Authors used video modality to classify the domain gender by face and then choose the appropriate speaker depending on the acoustic model. That combination provides substantial reduction of the word error rate (WER).

The deep neural network (DNN) requires significantly much training data when compared to classical HMM-GMM (Gaussian Mixture Models) techniques. It means that often training from scratch is not possible way. There exist techniques for improvement of ASR quality for particular groups of speakers. The DNN-HMM model has been adapted for a group with specific age and gender group for Italian language [14]. The baseline general model was trained there using several datasets that included adult male, adult female, and children utterances. New acoustic models were obtained by the continuation of training with the same parameters on each target subset. Each adapted model outperforms the baseline on its target test set (adult male, adult female, children). Thus, the authors showed that pretrained acoustic model can be used to adapt the neural network model for a certain gender group.

Transfer learning as an approach to adapting adult speech recognition system for children was successfully approbated in [15]. The next obvious way to train a neural network with lack of training data is finetuning. It starts new training for a new dataset but initialize the weights from pretrained network without substantial network modification. A new optimizer, hyperparameters, and output layer are used. This method of
model adaptation was used for accented speech from non-native English speakers and people with amyotrophic lateral sclerosis [16]. The authors of the latter paper trained two base models with different architectures on a large dataset. These pretrained models were used to improve accuracy for people with a non-standard speech by model tuning on a smaller dataset collected for a targeted group of people with amyotrophic lateral sclerosis and for non-native speakers. This work showed that WER reduction on target small datasets could be obtained by neural network adaptation via finetuning.

One more popular approach to adapt models for certain speaker domain is the usage of voice embeddings, e.g., i-vectors, x-vector, and d-vectors. The x-vector embeddings capture information and allow to build very accurate classifiers for speaker gender and speaker identification tasks [17] and outperforms i-vectors. The d-vectors showed its advantages for cocktail-party related tasks [18].

3 Materials and Methods

3.1 Data

**LibriSpeech.** is a common and popular dataset for speech recognition system experiments [23]. It includes transcribed ~960 hours of public domain audiobooks which are dictated by many speakers. Totally, for train purposes, the male sample of a LibriSpeech provides about 496 hours and the female is a bit above 465 hours. The validation part is about 8 hours as for males as females.

**L2 Arctic.** L2-Arctic corpus (v5.0) is a small dataset of non-native English speech [24]. It includes voice samples of 24 speakers (12 males and 12 females). One country dataset has 2 male and 2 female speakers. The total duration is 27.1 hours. We split the data into 90/10 and as in original paper [16]. We have chosen for validation dataset utterances without proper nouns.

3.2 Acoustic models

In this paper, we finetune the pretrained neural network for gender-based domain. We use the CTC-loss Transformer architecture, which provides competitive ASR results [19]. It is a sequence-to-sequence model, which is widely used in end-to-end speech processing. It consists of encoder and decoder parts and self-attention mechanism between. The encoder converts speech features into embeddings, self-attention learns the sequence information and provides it to the decoder. The last one could be considered as a small language model (LM), as it learns to predict characters or pieces according to language rules. The last experiments show that it could also show state-of-the-art result as a part of other models [20].

We prepare Transformer base models for further experiments. The model was pretrained with 960 hours of LibriSpeech dataset and we also prepare 2 models for each gender part of LibriSpeech. Thus, we have three base models for experiments.

The Librispeech-960 model was finetuned separately on both gender domains of LibriSpeech and L2 Arctic. We used ESPnet framework [21] that provides Pytorch implementation and pipeline for feature extraction inherited from Kaldi. Receipts for the
LibriSpeech dataset are supported from-the-box. Feature extraction for L2 Arctic we added by ourselves. We disabled RNN (recurrent neural network) language model decoding because it allows comparing acoustic models much objectively.

In our experiments with embeddings, we used pretrained 512 dimensioned voice x-vectors from the LibriSpeech Text-to-Speech pipeline in the ESPNet receipt [22]. We also modified the ASR feature extraction pipeline and incorporate x-vectors into the training pipeline. Two approaches have been implemented, namely, x-vector could be added to embeddings came from the encoder or concatenated with them.

3.3 Base model

We trained acoustic models during 100 epochs of LibriSpeech data. We used the default feature extraction pipeline for the LibriSpeech receipt in ESPnet. It assumes 80 filter banks as input for the acoustic model. To incorporate x-vectors we add MFCC (mel-frequency cepstral coefficients) extraction for each utterance. The model trained with a default for LibriSpeech receipt parameters: Noam optimizer, warmup step 25000, and learning rate 5.

Class scores (estimates of posterior probabilities) at the output of Transformer acoustic model can be fused with separate RNN-based language model. But we assume that experiments without the external language model provides more fair comparison and do not use LM in our experiments.

3.4 Finetuning

We changed various hyperparameters, optimizers, number of epochs to achieve the best result. We saved checkpoints for all 100 epochs during training of acoustic model, so that it is possible to start finetuning from any checkpoint. We conducted experiments with adaptation starting from the 20th, 50th, 60th, 80th, and 100th epochs. All experiments except beginning from 100th epochs lead to accuracy degradation. We tried to replace the Noam optimizer with Adadelta and Adam and play with its hyperparameters and the number of epochs. Our best result yielded with model tuning during 20 epochs with Adadelta, learning rate 0.1 and warmup step decreasing to 4000. All experiments were made. Our experiments with partial frozen layers confirmed that Transformer finetuning provides the best result if we do not freeze any layers similarly to other Encoder-Decoder network LAS.

4 Experimental results

4.1 LibriSpeech

We report detailed WER observing we measured for our three base models trained on LibriSpeech in Table 1. The general baseline model shows lower WER on all test sets than trained on gender subsets only. The main reason here is the size of the training data. The general model has been trained on almost twice larger dataset when compared to the training on gender specific data only.
Table 1. Average WER of base mode, LibriSpeech dataset.

|                      | test_clean | test_other | dev_clean | dev_other |
|----------------------|------------|------------|-----------|-----------|
|                      | M  | Fm | Full | M  | Fm | Full | M  | Fm | Full | M  | Fm | Full |
| Baseline (960h trained) | 4  | 4.6 | 4.3  | 11.2 | 9.3 | 10.2 | 4.4 | 3.8 | 4.1  | 11.2 | 9.3 | 10.2 |
| Male subset          | 7.4 | 9.8 | 8.6  | 17.4 | 18.4 | 17.9 | 8.9 | 8.3 | 8.6  | 17.0 | 18.8 | 17.8 |
| Female subset        | 8.7 | 7.9 | 8.3  | 20.0 | 14.7 | 17.4 | 9.5 | 6.9 | 8.2  | 20.5 | 14.6 | 17.5 |

Table 2. Average WER of finetuned model on male subset of LibriSpeech.

|                      | test_clean | test_other | dev_clean | dev_other |
|----------------------|------------|------------|-----------|-----------|
|                      | M  | Fm | Full | M  | Fm | Full | M  | Fm | Full | M  | Fm | Full |
| Baseline (960h trained) | 4  | 4.6 | 11.2 | 9.3 | 4.4 | 3.8 | 11.2 | 9.3 |
| Male adapted         | 3.8 | 4.6 | 11.0 | 9.4 | 4.3 | 3.7 | 11.0 | 9.5 |

The results of measurements for the finetuned model on the male subset of LibriSpeech is provided in Table 2. We established that gender-based fine-tuning provides some non-significant accuracy improvement ~5%, but also it provokes the same degradation for other groups. Taken into account the results of [14] where the pretrained model continues to be learned on the part of the dataset we expected much more improvement. This could be explained that our model is already well-trained and overfit on the same dataset is not able to bias the model. 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.

X-vector embeddings were examined in the next set of experiments. The training curve with the accuracy on the LibriSpeech validation dataset is shown in Fig. 1. Concatenation of x-vector and encoder embeddings gives results worse than the baseline. Hence, we conducted our measurement for the model, trained with summation of x-vectors. Table 3 shows that x-vector incorporation does not allow to improve accuracy significantly. Our adaptation experiments for the male subset does not show changes at all. Unfortunately, the model with voice embeddings is not suitable to improve the overall quality of gender adaptation. Here our expectation that we could increase the difference in accuracy between baseline and finetuned model is not confirmed, so that x-vectors seem to be useless in this task.

Table 3. Average WER of baseline trained with x-vector embedding, LibriSpeech dataset.

|                      | test_clean | test_other | dev_clean | dev_other |
|----------------------|------------|------------|-----------|-----------|
|                      |            |            |           |           |
| Baseline (960h trained) | 4.3 | 10.2 | 4.1 | 10.2 |
| Baseline with added X-vectors | 4.4 | **10.1** | 4.1 | 10.2 |
4.2 L2 Arctic

The next set of experiments is connected with L2 Arctic. We observed the changes of WER after tuning the base model on the test part of this dataset. Measurements of error for the L2 dev dataset and its gender subset are shown in Table 4. We achieved similar results as in the paper from Google. The accuracy is improved on the accented dataset after fine-tuning. But our model showed less error reduction when compared to the RNN-Transducer [16]. Our model is not so well trained and we do not use the language model because we did not set the goal to achieve the maximum WER reduction.

| Dataset          | Baseline | Finetuned (L2 Arctic train) |
|------------------|----------|-----------------------------|
| Arctic L2 dev (Full) | 29.5   | 19.1                        |
| Arctic L2 dev (M) | 27.7   | 20.1                        |
| Arctic L2 dev (Fm) | 31.2  | 18.4                        |

We expected that finetuning on gender subsets of L2 Arctic shows substantial improvement on separated test sets. But our experimental results from Table 5 show that adaptation on separated accented subsets yields just a tiny profit for females (~2%). As for the LibriSpeech dataset, gender female adaptation gives a bit better improvement than males.
Finally, we conducted an experiment with speakers from China. As for previous experiments, the WER is estimated for the base model but for utterances of male and female Chinese speakers only. Then we adapted acoustic model for male gender domains. The results are presented in Table 6. Here the adaptation for a group of Chinese males gives results better than for a group of all males from the dataset but still does not exceed the finetuning result on full L2 Arctic.

**Table 6.** Average WER of gender adaptation for Chinese male speakers.

|                | arctic_dev |     |     |
|----------------|------------|-----|-----|
|                | M          | Fm  |     |
| Baseline (960h trained) | 26.7       | 35.8|
| Male adapted   | **19.1**   | **33.9**|

## Conclusion

In this paper, we examined finetuning of DNN acoustic models to improve accuracy for male and female speaker by assuming that the gender of particular speaker is known. We trained the baseline Transformer model on LibriSpeech data and adapt it by male and female subsets of Librispeech. It was experimentally found that better result gives common practice like learning rate and warmup step decreasing and changing of optimizer from Noam to Adadelta. We obtained a small improvement (5%) when the baseline we bias on the male part of the dataset. However, incorporating x-vector between the encoder and decoder of the Transformer does not impact accuracy (Table 3). It leaves the same.

A series of experiments for L2 Arctic dataset with non-native English speech proved the accuracy improvements for personalizing models for accented speech. However, the adaptation of baseline on gender parts of L2 Arctic does not show significant accuracy improvement (1-2%). In fact, we do not obtain essential WER reduction by finetuning techniques for both datasets. In future, it is necessary to train other acoustic features because default Mel frequency filter banks do not represent the vocal differences between females and males. We should try to add pitch features or replace filter banks with MFCCs. Moreover, it is necessary to repeat our experiments with other architecture because it could be more sensitive for finetuning.
References

1. Klatt, Dennis H., Laura C. Klatt.: "Analysis, synthesis, and perception of voice quality variations among female and male talkers." the Journal of the Acoustical Society of America 87.2, 820-857 (1990).
2. Mendoza, Elvira, et al.: "Differences in voice quality between men and women: use of the long-term average spectrum (LTAS)." Journal of voice 10.1, 59-66 (1996).
3. Savchenko, Andrey V.: "Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output ConvNet." PeerJ Computer Science 5: e197 (2019).
4. Do, Trung Dung, Van Huan Nguyen, Hakil Kim.: "Real-time and robust multiple-view gender classification using gait features in video surveillance." Pattern Analysis and Applications 23.1: 399-413 (2020).
5. Son, Guiyoung, Soonil Kwon, Neungsoo Park: "Gender classification based on the non-lexical cues of emergency calls with recurrent neural networks (RNN)." Symmetry 11.4: 525 (2019).
6. Markitantov, Maxim, Oxana Verkholyak: "Automatic Recognition of Speaker Age and Gender Based on Deep Neural Networks." International Conference on Speech and Computer. Springer, Cham, (2019).
7. Pépiot, Erwan.: "Voice, speech and gender: male-female acoustic differences and cross-language variation in English and French speakers." Corela. Cognition, représentation, langage HS-16 (2015).
8. Zäske, Romi, et al.: "Perceiving vocal age and gender: An adaptation approach." Acta psychologica 144.3 583-593, (2013).
9. Hubbard, Daniel J., Peter F. Assmann. "Perceptual adaptation to gender and expressive properties in speech: The role of fundamental frequency." The Journal of the Acoustical Society of America 133.4 (2013): 2367-2376.
10. Hosseini-Asl, Ehsan, et al.: "A multi-discriminator cycleGAN for unsupervised non-parallel speech domain adaptation." arXiv preprint arXiv:1804.00522 (2018).
11. Ramya, T., et al.: "Analysis on MAP and MLLR based speaker adaptation techniques in speech recognition." 2014 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2014]. IEEE, (2014).
12. Savchenko, Andrey V.: "Semi-automated Speaker Adaptation: How to Control the Quality of Adaptation?." International Conference on Image and Signal Processing. Springer, Cham, (2014).
13. Campr, Pavel, et al.: "Online speaker adaptation of an acoustic model using face recognition." International Conference on Text, Speech and Dialogue. Springer, Berlin, Heidelberg, (2013).
14. Serizel, Romain, Diego Giuliani: "Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children." Natural Language Engineering 23.3: 325-350 (2017).
15. Shivakumar, Prashanth Gurunath, and Panayiotis Georgiou. "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations." arXiv preprint arXiv: 1805.03322 (2018).
16. Shor, Joel, et al.: "Personalizing ASR for dysarthric and accented speech with limited data." arXiv preprint arXiv:1907.13511 (2019).
17. Raj, Desh, et al.: "Probing the information encoded in x-vectors." 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, (2019).
18. Wang, Quan, et al. "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking." arXiv preprint arXiv:1810.04826 (2018).

19. Karita, Shigeki, et al.: "A comparative study on transformer vs RNN in speech applications." IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, (2019).

20. Zhang, Qian, et al.: "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, (2020).

21. Watanabe, Shinji, et al. "Espnet: End-to-end speech processing toolkit." arXiv preprint arXiv:1804.00015 (2018).

22. Hayashi, Tomoki, et al.: "ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, (2020).

23. Panayotov, Vassil, et al: "LibriSpeech: an ASR corpus based on public domain audio books." 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, (2015).

24. Zhao, Guanlong, et al.: "L2-ARCTIC: A non-native English speech corpus." Perception Sensing Instrumentation Lab (2018).