INVESTIGATION INTO JOINT OPTIMIZATION OF SINGLE CHANNEL SPEECH ENHANCEMENT AND ACOUSTIC MODELING FOR ROBUST ASR

Tobias Menne, Ralf Schlüter, Hermann Ney

Human Language Technology and Pattern Recognition, Computer Science Department,
RWTH Aachen University, Aachen, Germany

ABSTRACT

This paper investigates the joint optimization of single channel speech enhancement and the acoustic model of a hybrid DNN-HMM system for noise robust ASR. Two enhancement methods are investigated. A masking of the noisy speech signal with a speech mask estimated by a DNN based mask estimator, as well as a parametric Wiener filter employing a DNN based noise estimator and a DNN based frame wise estimation of the filter parameters. Those components are jointly optimized with the acoustic model of the ASR system. It is shown that the Wiener filter approach can be used to improve the performance of a state-of-the-art single-channel ASR system on the single channel track of the CHiME-4 data, where the WER of the real ealuation set is reduced from $11.6\,\%$ to $10.5\,\%$.

Index Terms— robust ASR, single-channel ASR, joint training, speech enhancement, CHiME-4

1. INTRODUCTION

Significant improvements in the performance of automatic speech recognition (ASR) have been achieved over the last decade. The performance gains are driven especially by the application of deep learning techniques [1]. Nevertheless performance in noisy scenarios is still significantly worse than the performance on undisturbed speech [2, 3]. To mitigate this effect, multiple microphones and multi-channel speech enhancement techniques like filter-and-sum beamforming are often used to improve the signal quality and in turn ASR performance as seen e.g. in the submissions to the 4^{th} CHiME challenge [4, 5, 6, 7].

More recently, approaches that jointly optimize multi-channel speech enhancement algorithms and the acoustic model to improve ASR performance have been proposed. To this end statistically optimal beamformers, which are supported by data driven parameterizable mask estimators are jointly trained to minimize the acoustic model training criterion [8, 9, 10]. That work was mainly focused on multi-channel approaches.

Single channel speech enhancement techniques have proven to be less effective and often improvements in speech quality metrics like perceptual evaluation of speech quality (PESQ) do not translate into better ASR performance [5]. Thus the ASR performance is still significantly worse when only a single microphone is available [4, 5, 6].

In related work as e.g. [11] word error rate (WER) improvements on simulated noisy data are reported by utilizing deep neural networks (DNNs) as regression models to enhance noisy speech and training them jointly with the acoustic model of an hybrid ASR system. In [5] a DNN is used to estimate speech masks, which are directly applied to the noisy signal for enhancement. No improvements could be reported with this method on the real CHiME-4 data, but no joint training of the speech enhancement and acoustic model has been done in [5]. In this work we investigate the joint optimization of single channel speech enhancement and acoustic modeling, while focusing on the approach of utilizing DNNs as mask estimating networks. Additionally to using the direct masking approach used in [5], we investigate the utilization of a parametric Wiener filter. The Wiener filter has been investigated as a preprocessing approach for ASR before with mixed results [12, 13, 14]. Here we focus on it's integration with the hybrid acoustic model and utilization of a DNN based mask estimate for the noise estimation within the filtering. This turns out to be the key to WER improvements in our experiments. The rest of the paper is organized as follows. An overview over the investigated mask based single channel speech enhancement methods is given in Section 2. The experimental setup including the ASR system is discussed in Section 3 and the results are presented and discussed in Section 4.

2. SPEECH ENHANCEMENT

This work investigates an integration of speech enhancement and acoustic model. The short-time Fourier transform (STFT) domain input $Y_{t,f}$ is obtained from a noisy recording containing a speech component $S_{t,f}$ and a noise component $N_{t,f}$

$$Y_{t,f} = S_{t,f} + N_{t,f} \tag{1}$$

where t is the time frame index out of T total frames and f is the frequency bin index out of F total frequency bins. The speech enhancement module estimates the clean signal

$$\hat{S}_{t,f} = \mathrm{f}_{\Theta_\mathrm{f}}(\mathbf{Y}, t, f) \tag{2}$$

where Θ_f are trainable parameters of the speech enhancement. Acoustic features $x_t = \mathrm{g}(\hat{\mathbf{S}}, t)$ are extracted from the enhanced signal and used as input features for the DNN hybrid acoustic model returning the posterior likelihoods $\mathrm{p}^{\mathrm{AM}}_{\Theta_{\mathrm{AM}}}(s_t|x_1^T)$ for state s at time t given the feature vector

sequence x_1^T. Where Θ_{AM} are the trainable parameters of the acoustic model. Through the integration of speech enhancement and acoustic model into a single model, Θ_{f} and Θ_{AM} can be jointly optimized towards an ASR loss function.

The individual building blocks of the system are described in the following.

2.1. Estimation of speech and noise masks

This work employs a mask estimating neural network similar to the ones described in [15, 10], where the output at time frame t of the network is

$$[\lambda_{t,1}^{(S)}, \dots \lambda_{t,F}^{(S)}, \lambda_{t,1}^{(N)}, \dots, \lambda_{t,F}^{(N)}] = \mathrm{h}_{\Theta_\lambda}^{(\lambda)}(\mathbf{Y}, t) \tag{3}$$

Where Θ_λ are the trainable parameters of the mask estimator network. The output values $\lambda_{t,f}^{(\nu)}$ are restricted to the value range $(0, 1)$ by using a sigmoid output layer. The output $\lambda_{t,f}^{(\nu)}$ for $\nu \in \{S, N\}$ is interpreted as the likelihood of speech or noise being present at time-frequency bin (t, f), respectively.

2.2. Direct masking

The approach of directly applying a speech mask $\lambda_{t,f}^{(S)}$ to the noisy signal $Y_{t,f}$ to enhance the speech, which has been investigated as a separate preprocessing in [5] is jointly optimized with the acoustic model in this work. The enhanced signal is computed as

$$\hat{S}_{t,f} = \mathrm{f}_{\Theta_{\mathrm{DM}}}^{(\mathrm{DM})}(\mathbf{Y}, t, f) = \lambda_{t,f}^{(S)} \cdot Y_{t,f} \tag{4}$$

The trainable parameters of the direct masking (DM) are thus the parameters of the mask estimator $\Theta_{\mathrm{DM}} = \{\Theta_\lambda\}$

2.3. Parametric Wiener filter

Direct masking of a speech signal often introduces a strong degree of distortion to the enhanced speech signal. To mitigate this effect, the parametric Wiener filter (PW) as described in [16, 17] offers more flexibility in controlling the trade off between speech distortion and noise suppression. The output of the Wiener filter is computed as

$$\hat{S}_{t,f} = \mathrm{f}_{\Theta_{\mathrm{PW}}}^{\mathrm{PW}}(\mathbf{Y}, t, f) = \left| \frac{|Y_{t,f}|^p - l \cdot |\hat{N}_{t,f}|^p}{|Y_{t,f}|^p} \right|^{\frac{1}{q}} \cdot Y_{t,f} \tag{5}$$

where $\hat{N}_{t,f}$ is an estimate of the noise power spectrum and l, p, q are the parameters that control the parametric Wiener filter and will be referred to as Wiener parameters in the following. Usually those parameters are set manually to control the trade off between noise suppression and speech distortion. In this work a neural network is used to obtain frame wise estimates of those parameters.

$$[l_t, p_t, q_t] = \mathrm{h}_{\Theta_{\mathrm{PE}}}^{(\mathrm{PE})}(\mathbf{Y}, t) \tag{6}$$

where Θ_{PE} are the trainable parameters of the Wiener parameter estimation network.

Furthermore a noise estimate

$$\hat{N}_{t,f} = \mathrm{h}_{\Theta_{\mathrm{NE}}}^{(\mathrm{NE})}(\mathbf{Y}, t, f) \tag{7}$$

is required for the Wiener filter, where Θ_{NE} are potential trainable parameters of the noise estimator. Thus $\Theta_{\mathrm{PW}} = \{\Theta_{\mathrm{NE}}, \Theta_{\mathrm{PE}}\}$ are the trainable parameters of the Wiener filter used here.

Two different noise estimators are investigated in this work. The first one is a frame average of the first P frames

$$\hat{N}_{t,f} = \mathrm{h}^{(\mathrm{NE},1)}(\mathbf{Y}, f) = \frac{1}{P} \cdot \sum_{t'=1}^{P} Y_{t',f} \tag{8}$$

obtaining a static noise estimate for the whole signal. The second noise estimator is based on a noise mask estimate $\lambda_{t,f}^{(N)}$ and the noise estimate is obtained by masking the input signal

$$\hat{N}_{t,f} = \mathrm{h}_{\Theta_{\mathrm{NE},2}}^{(\mathrm{NE},2)}(\mathbf{Y}, t, f) = \lambda_{t,f}^{(N)} \cdot Y_{t,f} \tag{9}$$

The noise mask $\lambda_{t,f}^{(N)}$ is obtained as described in 2.1.

3. EXPERIMENTAL SETUP

The proposed joint training of the single channel speech enhancement and the acoustic model is evaluated on the single channel track of the CHiME-4 speech-recognition task [7]. The CHiME-4 dataset contains simulated and real $16\,\mathrm{kHz}$ data which was recorded, with a hand held device, in four different noise environments. This work is focused on the real data.

80 dimensional log-mel filterbank features are used as input features for the acoustic model, where the STFT is using a hanning window applied to a $25\,\mathrm{ms}$ frame with a frame shift of $10\,\mathrm{ms}$.

The input features are unnormalized but a 80 dimensional linear input layer employing batch normalization [18] is used as a first layer of the acoustic model. The linear layer is followed by 5 bidirectional long short-term memory (BLSTM) layers with 600 units each. The output is a softmax layer with 1501 units. The acoustic model is pretrained with the cross entropy (CE) loss function in the same manner described in [4] using the unprocessed data as input signals.

The mask estimation network, described in Section 2.1, consists of a BLSTM layer with 256 units followed by two fully connected layers with 512 units and ReLU activation function. The mask estimation output layer is a fully connected layer with sigmoid activation function. The input of the mask estimation network is the magnitude spectrum of the noisy input signal using the same configuration for the STFT as for the feature extraction. The mask estimation network is initialized by a separate pretraining similar to the one described in [15], where the simulated training data is used to compute ideal binary masks as training targets.

The parameter estimation network, described in Section 2.3, consists of a single BLSTM layer with 1024 units followed by a fully connected layer with sigmoid activation function and 3 output units. Thus the parameters l, p and q are restricted to a value range of $(0, 1)$. In preliminary experiments we investigated variations of the activation function

of the output layer, as e.g. $2 \cdot \text{sigmoid}$, $10 \cdot \text{sigmoid}$ and the identity function, since a restriction of the parameters to the value range of $(0, 1)$ is not generally required. The results indicated that reasonable results could be obtained with other activation functions as well, but the simple sigmoid output function offered a good trade off between performance and stability of training convergence towards smaller changes in hyper parameters. The parameter estimation network is always trained from scratch during the joint training.

The joint training is done for 2 epochs with the CE loss function on the complete training set of the CHiME-4 data. For every combination of speech enhancement with the acoustic model, the hyperparameters of the joint training like learning rate, gradient noise and dropout are tuned separately on the development set. It is noteworthy, that the direct masking approach was tuned with the same effort as the Wiener filter approach.

Decoding is done with a 5-gram language model. In a post processing step a recurrent neural network (RNN) language model lattice rescoring is done. The RNN language model is a 3 layer long short-term memory (LSTM) and is further described in [4].

4. EXPERIMENTAL RESULTS & DISCUSSION

Table 1 shows the WER for the different combinations of single channel speech enhancement with the acoustic model. The authors of [5] found that ASR performance declined, when using the direct masking approach even when the processed data was seen during training by the acoustic model. Consistent with their findings, we also could not obtain performance gains with the direct masking approach, even when jointly optimizing the mask estimation network and the acoustic model. The parametric Wiener filter approach on the other hand shows significant performance gains after joint optimization, especially when used with the mask based noise estimation.

Table 1. Average WER (%) for various speech enhancement methods after joint optimization.

System		Dev	Eval
Front-end	Noise estimator		
-		6.6	11.6
DM		7.0	13.4
PW	frame avg.	6.2	11.3
	masking	**6.0**	**10.6**

Figure 1 shows the frame wise estimated parameters of the Wiener filter for an example signal. The plots show, that the variation of the parameter values over the signal duration is relatively small especially when using the frame average as a noise estimate.

Figure 2 shows the distribution of the frame wise parameters values over the development set. The histograms show, that the parameter values are distributed in a relatively narrow value range, but the specific parameter value ranges differ for the different noise estimators. The WERs in Table 2 confirm, that the frame wise estimation of the parameters is not essential for the performance gain and that using fixed values

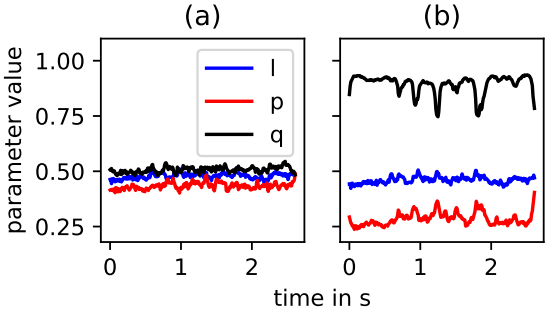

Fig. 1. Example of frame wise parameter estimation of l, p and q of Equation 5 for the signal F04_053C0108_STR for (a) using the first-T frames noise estimator (b) using the mask based noise estimator

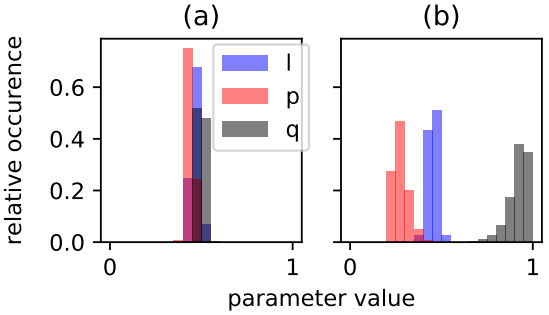

Fig. 2. Histogram of the parameter values of l, p and q from Equation 5 computed on the development set after joint training. (a) using the first-T frames noise estimator (b) using the mask based noise estimator

provided from the jointly optimized model as average over the development set works equally well.

Table 2. Average WER (%) for the parametric Wiener approach after joint optimization. The Wiener parameters are either estimated separately for each frame (frame wise), averaged over the complete signal (seq. avg.) or fixed for all signals to the average computed over the development set (dev avg.).

System			Dev	Eval
Front-end	Noise estimator	Wiener param.		
-			6.6	11.6
PW	frame avg.	frame wise	6.2	11.3
		seq. avg.	6.2	11.4
		dev avg.	6.2	11.4
	masking	frame wise	6.0	10.6
		seq. avg.	6.0	10.5
		dev avg.	5.9	10.6

Table 3 shows an extreme performance drop if the direct masking enhancement is used without the joint optimization. This is consistent with the observation in [5], that the acoustic model needs to have seen signals processed by the

direct masking enhancement method during training to obtain reasonable ASR performance with this enhancement approach. Also the parametric Wiener approach utilizing the frame average as noise estimate benefits from the adaptation of the acoustic model during joint optimization. In contrast to this the parametric Wiener filter utilizing the mask based noise estimate works equally well with the original acoustic model and mask estimator if the Wiener parameter configuration would be known before the joint optimization. Table 4 on the other hand indicates, that knowing the correct parameter configuration for the specific noise estimator is a key factor to obtain improved performance with this approach. Thus a key benefit of the approach presented here is the derivation of the Wiener parameter configuration during the joint optimization of the integrated model. Furthermore the results indicate, that whether or not an adaptation of the acoustic model parameters Θ_{AM} and noise estimator parameters Θ_{NE} is beneficial when using the parametric Wiener filter, depends on the noise estimator. Thus the proposed method of deriving the Wiener parameters during joint training is preferable over a simple grid search during decoding.

Table 3. Average WER (%) for various speech enhancement methods before and after joint optimization. Fixed values are used for the Wiener parameters. Those parameters have been derived from the respective jointly optimized model averaged over the development set.

System			Dev	Eval
Front-end	Noise estimator	Joint training		
-			6.6	11.6
DM		-	29.3	32.1
		×	7.0	13.4
PW	frame avg.	-	7.6	13.9
		×	6.2	11.4
	masking	-	**5.9**	**10.5**
		×	**5.9**	10.6

Table 4. Average WER (%) for the parametric Wiener approach with the original acoustic model and mask estimator network. The Wiener parameters values are fixed and computed as averages over the development set using the jointly trained models either using the frame averaged or mask based noise estimate.

Front-end	Wiener param.	Dev	Eval
-		6.6	11.6
PW	masking	5.9	10.5
	frame avg.	6.9	13.0

Figure 3 shows the enhanced speech after the different processing methods. The plots indicate, that the joint optimization of the mask estimator and the acoustic model in the direct masking approach moves the mask estimator towards a weaker suppression of the time-frequency bins without harmonics.

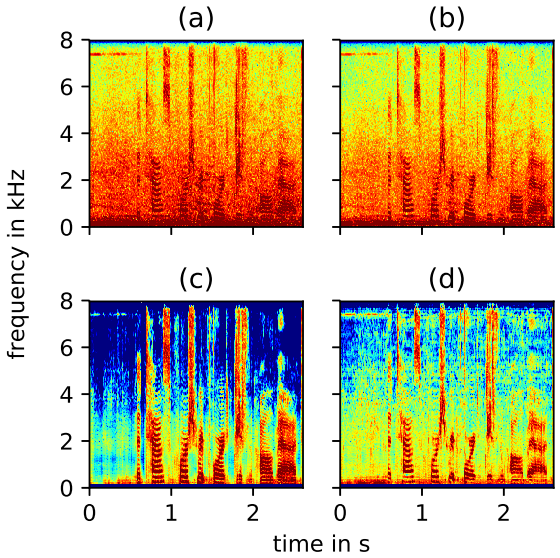

Fig. 3. Spectrograms of F04_053C0108_STR ("The task force report isn't all bad") after various processings. (a) original signal (b) Wiener filter of jointly trained model with mask based noise estimation (c) direct masking before joint training (d) direct masking after joint training

5. CONCLUSION

This paper describes the integration of single channel speech enhancement and acoustic modeling for joint parameter optimization. In addition to the direct masking approach e.g. used in [5], the utilization of a parametric Wiener filter was introduced and tested. The performance was evaluated on real noisy data of the single channel track of the CHiME-4 data. By utilization of the parametric Wiener filter the WER of the real evaluation set of the CHiME-4 data was reduced from $11.6\,\%$ to $10.5\,\%$ which shows, that the parametric Wiener filter in combination with a powerful DNN based noise mask estimator can be beneficial for single channel noise robust ASR. Our future work will explore approaches to avoid pretraining of the DNN based mask estimator, employed in the parametric Wiener filter, and to instead train it from scratch during the joint optimization. The integration of the speech enhancement and the acoustic modeling presented in this work allows for utilizing label feedback to achieve this.

6. ACKNOWLEDGMENTS

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement No. 694537 and under the Marie Skłodowska-Curie grant agreement No. 644283. The work reflects only the authors' views and none of the funding agencies is responsible for any use that may be made of the information it contains.

7. REFERENCES

[1] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Jaitly Navdeep, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[2] Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, Apr 2014.

[3] Marc Delcroix, Takuya Yoshioka, Atsunori Ogawa, Yotaro Kubo, Masakiyo Fujimoto, Nobutaka Ito, Keisuke Kinoshita, Miquel Espi, Shoko Araki, Hori Takaaki, and Tomohiro Nakatani, "Strategies for distant speech recognitionin reverberant environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 4, pp. 60–74, Jul 2015.

[4] Tobias Menne, Jahn Heymann, Anastasios Alexandridis, Kazuki Irie, Albert Zeyer, Markus Kitza, Pavel Golik, Ilia Kulikov, Lukas Drude, Ralf Schlüter, Hermann Ney, Reinhold Haeb-Umbach, and Athanasios Mouchtaris, "The RWTH/UPB/FORTH system combination for the 4th CHiME challenge evaluation," in *Proc. of the 4th Intl. Workshop on Speech Processing in Everyday Environments (CHiME 2016)*, San Francisco, CA, Sept. 2016, pp. 39–44.

[5] Szu-Jui Chen, Aswin Shanmugam Subramanian, Hainan Xu, and Shinji Watanabe, "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," in *Proc. Interspeech*, Hyderabad, India, Sept. 2018, pp. 1571–1575.

[6] Yan-Hui Tu, Jun Du, Lei Sun, Feng Ma, and Chin-Hui Lee, "On design of robust deep models for CHiME-4 multi-channel speech recognition with multiple configurations of array microphones," in *Proc. Interspeech*, Stockholm, Sweden, Aug 2017, pp. 394–398.

[7] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, Nov 2017.

[8] Jahn Heymann, Lukas Durde, Christoph Boeddeker, Patrick Hanebrink, and Reinhold Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, Mar. 2017, pp. 5325–5329.

[9] Christoph Boeddeker, Patrick Hanebrink, Lukas Durde, Jahn Heymann, and Reinhold Haeb-Umbach, "Optimizing neural-network supported acoustic beamforming by algorithmic differentiation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, Mar. 2017, pp. 171–175.

[10] Tobias Menne, Ralf Schlüter, and Hermann Ney, "Speaker adapted beamforming for multi-channel automatic speech recognition," in *accepted for publication in Spoken Language Technology Workshop (SLT)*, Athens, Greece, Dec. 2018.

[11] Tian Gao, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4375–4379.

[12] Yifan Gong, "Speech recognition in noisy environments: A survey," *Speech communication*, vol. 16, no. 3, pp. 261–291, 1995.

[13] Jian Wu, Jasha Droppo, Li Deng, and Alex Acero, "A noise-robust asr front-end using wiener filter constructed from mmse estimation of clean speech and noise," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*. IEEE, 2003, pp. 321–326.

[14] Wang Xu, Yonghui Guo, Bingxi Wang, Xingbing Wang, and Zhifei Mai, "A noise robust front-end using wiener filter, probability model and cms for asr," in *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on*. IEEE, 2005, pp. 102–105.

[15] Jahn Heymann, Lukas Drude, Aleksej Chinaev, and Reinhold Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, AZ, Dec. 2015, pp. 444–451.

[16] Jacob Benesty, M Mohan Sondhi, and Yiteng Huang, *Springer handbook of speech processing*, Springer-Verlag Berlin Heidelberg, 2008.

[17] Jae S. Lim and Alan V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[18] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.