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 evaluation 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 4th 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 $N_{t,f}$

$$Y_{t,f} = S_{t,f} + N_{t,f} \quad (1)$$

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

$$\hat{S}_{t,f} = f_{\Theta_{t}} (Y, t, f) \quad (2)$$

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

©2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses. DOI: 10.1109/ICASSP.2019.8682460 Publisher version: https://ieeexplore.ieee.org/document/8682460
The sequence $x_t^T$. Where $\Theta_{AM}$ are the trainable parameters of the acoustic model. Through the integration of speech enhancement and acoustic model into a single model, $\Theta_t$ and $\Theta_{AM}$ 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

$$[\lambda_{t,1}^{(S)}, \ldots, \lambda_{t,F}^{(S)}, \lambda_{t,1}^{(N)}, \ldots, \lambda_{t,F}^{(N)}] = h^{(\lambda)}(Y, t)$$

(3)

Where $\Theta_{\lambda}$ are the trainable parameters of the mask estimator network. The output values $\lambda_{t,f}^{(\mu)}$ 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

$$\hat{S}_{t,f} = f^{(DM)}(Y, t, f) = \lambda_{t,f}^{(S)} \cdot Y_{t,f}$$

(4)

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

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

$$\hat{S}_{t,f} = f^{PW}(Y, t, f) = \frac{|Y_{t,f}|^p - l \cdot |\hat{N}_{t,f}|^p}{|Y_{t,f}|^p} \cdot Y_{t,f}$$

(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.

$$[l_t, p_t, q_t] = h^{(PE)}(Y, t)$$

(6)

where $\Theta_{PE}$ are the trainable parameters of the Wiener parameter estimation network.

Furthermore a noise estimate

$$\hat{N}_{t,f} = h^{(NE)}(Y, t, f)$$

(7)

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

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

(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

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

(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 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 ms frame with a frame shift of 10 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 5 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 · sigmoid, 10 · 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.

<table>
<thead>
<tr>
<th>System</th>
<th>Dev</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Front-end</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noise estimator</td>
<td>6.6</td>
<td>11.6</td>
</tr>
<tr>
<td>DM</td>
<td>7.0</td>
<td>13.4</td>
</tr>
<tr>
<td>PW</td>
<td>6.2</td>
<td>11.3</td>
</tr>
<tr>
<td>masking</td>
<td>6.0</td>
<td>10.6</td>
</tr>
</tbody>
</table>

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 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.).

<table>
<thead>
<tr>
<th>System</th>
<th>Dev</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Front-end</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noise estimator</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wiener param.</td>
<td>6.6</td>
<td>11.6</td>
</tr>
<tr>
<td>PW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>frame avg.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>seq. avg.</td>
<td>6.2</td>
<td>11.3</td>
</tr>
<tr>
<td>dev. avg.</td>
<td>6.2</td>
<td>11.4</td>
</tr>
<tr>
<td>Frame wise</td>
<td>6.0</td>
<td>10.6</td>
</tr>
<tr>
<td>seq. avg.</td>
<td>6.0</td>
<td>10.5</td>
</tr>
<tr>
<td>dev. avg.</td>
<td>5.9</td>
<td>10.6</td>
</tr>
</tbody>
</table>

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 $\Theta_{AM}$ and noise estimator parameters $\Theta_{NE}$ 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.

<table>
<thead>
<tr>
<th>System</th>
<th>Front-end</th>
<th>Noise estimator</th>
<th>Joint training</th>
<th>Dev</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>-</td>
<td>-</td>
<td>×</td>
<td>6.6</td>
<td>11.6</td>
</tr>
<tr>
<td>PW</td>
<td>-</td>
<td>frame avg.</td>
<td>×</td>
<td>29.3</td>
<td>32.1</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>masking</td>
<td>×</td>
<td>7.6</td>
<td>13.9</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>×</td>
<td></td>
<td>6.2</td>
<td>11.4</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>System</th>
<th>Front-end</th>
<th>Wiener param.</th>
<th>Dev</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>-</td>
<td>6.6</td>
<td>11.6</td>
<td></td>
</tr>
<tr>
<td>PW</td>
<td>masking</td>
<td>5.9</td>
<td>10.5</td>
<td></td>
</tr>
<tr>
<td>PW</td>
<td>frame avg.</td>
<td>6.9</td>
<td>13.0</td>
<td></td>
</tr>
</tbody>
</table>

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.

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


