Learning Noise Invariant Features Through Transfer Learning for Robust End-to-End Speech Recognition

Citation for published version:

Digital Object Identifier (DOI):
10.1109/ICASSP40776.2020.9053169

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
LEARNING NOISE INVARIANT FEATURES THROUGH TRANSFER LEARNING FOR ROBUST END-TO-END SPEECH RECOGNITION

Shucong Zhang 1, Cong-Thanh Do 2, Rama Doddipatla 2 and Steve Renals 1

1 Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
2 Toshiba Cambridge Research Laboratory, Cambridge, United Kingdom

ABSTRACT
End-to-end models yield impressive speech recognition results on clean datasets while having inferior performance on noisy datasets. To address this, we propose transfer learning from a clean dataset (WSJ) to a noisy dataset (CHiME-4) for connectionist temporal classification models. We argue that the clean classifier (the upper layers of a neural network trained on clean data) can force the feature extractor (the lower layers) to learn the underlying noise invariant patterns in the noisy dataset. While training on the noisy dataset, the clean classifier is either frozen or trained with a small learning rate. The feature extractor is trained with no learning rate re-scaling. The proposed method gives up to 15.5% relative character error rate (CER) reduction compared to models trained only on CHiME-4. Furthermore, we use the test sets of Aurora-4 to perform evaluation on unseen noisy conditions. Our method has significantly lower CERs (11.3% relative on average) on all 14 Aurora-4 test sets compared to the conventional transfer learning method (no learning rate re-scale for any layer), indicating our method enables the model to learn noise invariant features.

Index Terms— end-to-end, robust speech recognition, transfer learning

1. INTRODUCTION
End-to-end speech recognition models simplify the training procedure compared to conventional hybrid systems and have offered impressive performance [1–3]. However, end-to-end models usually have inferior results on noisy data [4, 5].

Several methods have been proposed to help training on noisy data by exploiting clean data, such as teacher-student learning [6–8] and multi-task learning [9–11]. However, in these methods, typically require parallel clean/noisy data, which limits their usefulness. Transfer learning transfers the knowledge learned from the source domain to the target domain and does not require parallel data. In this work, to exploit non-parallel clean and noisy data in training end-to-end speech recognition models, we propose a novel transfer learning from clean speech data to noisy speech data.

Transfer learning has been widely employed in training speech recognition models for low-resource languages [12–17]. The low-level features of different languages are generally similar. Thus, a model is usually trained on a well-resourced language and the feature extractor (the lower layers of a neural network) is transferred across languages. However, for transfer learning from clean to noisy data, although the underlying patterns should be invariant to the noise conditions, it is not suitable to transfer the feature extractor owing to the mismatch of acoustic conditions.

We propose to transfer the classifier (the upper layers) from clean to noisy data, rather than transferring the feature extractor, by first training the classifier on the clean dataset. While training on the noisy dataset, the clean classifier is either frozen or tuned using a small learning rate; the feature extractor is trained using the normal learning rate. The feature extractor is constrained to learn features that match the clean classifier. Since the features fit the clean classifier reflect more about the underlying patterns, the clean classifier helps the feature extractor to learn the noise invariant patterns from the noisy data.

We apply the proposed transfer learning method to train connectionist temporal classification (CTC) models transferring from WSJ [18] to CHiME-4 [19, 20]. Compared to models trained directly on CHiME-4, the models trained with the proposed method reduce the character error rate (CER) by up to 15.5% relative. We also tested the performance of our method on unseen noise conditions using the Aurora-4 test sets [21]. Compared to conventional transfer learning in which there is no re-scale of the learning rate for any layer, the proposed method has significantly lower CERs on all 14 test sets with an average relative reduction in CER of 11%. These experiments indicate that the proposed transfer learning method helps the model to learn noise invariant features.

2. RELATED WORK
The core idea of our proposed transfer learning approach is that clean speech features should be similar to features which are invariant to different noise conditions. Thus, making the features extracted from noisy data similar to features extracted from clean data should be helpful.
To achieve this objective, teacher-student learning is applied [6–8]. In general, parallel clean/noisy data is required for teacher-student, with the noisy data often generated by adding noise to the clean data. The teacher model is trained on the clean dataset, and the output distribution of the teacher model of clean utterances is used as soft labels for the parallel noisy utterances. When training the student model on the noisy utterances, we employ an objective which minimizes the KL divergence between the soft labels and the output distribution of the student model. The transcript of the utterance is viewed as a hard label sequence and may be optionally used when training the student model.

Multi-task learning also helps in training robust speech recognition models by exploiting parallel data [9–11]. In multi-task learning, the main training objective is to recognize the noisy utterance and the secondary objective can be to reconstruct the clean utterance. An alternative secondary objective is to minimize the distance of the output of each layer between the model training on the noisy utterance and the model training on the parallel clean utterance.

Although these methods enable models to learn domain invariant features and improve the speech recognition results on noisy data, parallel data is necessary. In contrast, our proposed transfer learning method exploits knowledge learned on non-parallel clean data.

Transfer learning for low-resource languages [12–17] uses feature extractors trained on well-resourced languages, following which the classifier is reinitialized and retrained using the low-resource language. The feature extractor is either jointly trained or kept frozen, following an optional fine-tuning stage. For our proposed method, we transfer the classifier to force the feature extractor to learn noise invariant features.

### 3. CONNECTIONIST TEMPORAL CLASSIFICATION

Connectionist temporal classification (CTC) models [22] belong to the family of sequence-to-sequence models. They can be applied to end-to-end speech recognition since this model is alignment-free – it considers all the valid alignments. In this work, the inputs for the CTC models are acoustic features and the outputs are characters.

For an input sequence \( X = x_1, \ldots, x_t \), a valid output sequence \( A = a'_1, \ldots, a'_t \) contains repeated characters and blank symbols (−). The repeated characters between blank symbols will be merged into one character to generate the true output sequence \( Y = y_1, \ldots, y_s \). For example, for \( X = x_1, \ldots, x_5 \), \( A = c, c, -, a, t \) is a valid output for \( Y = c, a, t \). During training, the model maximizes the probability of the ground truth character sequence, which is the sum of the probability of all valid alignments:

\[
P(Y|X) = \sum_{A \in S} P(A|X),
\]

where \( S \) represents the set of valid alignments. CTC models are usually using by bidirectional recurrent neural networks, and the probability of each valid sequence is given by

\[
P(A|X) = \prod_{i=1}^{t} P(a_i|X).
\]

The probability \( P(Y|X) \) can be computed efficiently through a forward-backward algorithm. The decoding can be performed in a greedy way or using beam search.

### 4. TRANSFER LEARNING

We view the top layers of a model as a classifier and the layers beneath these top layers as a feature extractor. The classifier divides the space, while the feature extractor provides features based on this partition of the space. A well-trained clean model (i.e., a model which has a good speech recognition performance on the clean dataset) gives a “clean classifier”. We assume the features extracted by the well-trained clean model are close to the underlying patterns. Moreover, these features match the clean classifier. If we transfer the clean classifier and train the feature extractor on the noisy data, then the feature extractor is forced to extract features that fit the clean classifier. Since the features that fit the clean classifier should be close to the underlying patterns, the feature extractor is forced to extract noise invariant features from the noisy data.

In our transfer learning method, we do not only consider the output softmax layer as the classifier. We also view the grouping of the softmax layer and several other upper layers as the classifier. With more layers, the classifier is more powerful and may ease the burden of learning features from noisy data for the feature extractor. However, more layers for the classifier also means fewer layers for the feature extractor. It also makes the feature extractor less powerful and less flexible. Thus, with too few layers, the feature extractor may not have the capacity to fit the classifier. In our experiments, we set the number of layers for the classifier using the validation set.

While training on the noisy data, the clean classifier is either frozen or tuned with a small learning rate. The feature extractor is either reinitialized or initialized with the weights of the clean feature extractor, as if the training on the clean data is considered as a pre-training stage. The feature extractor is trained with the normal learning rate without any learning rate re-scaling.

### 5. EXPERIMENTS

We apply the proposed transfer learning method from WSJ [18] to CHiME-4 [19, 20]. For WSJ, we use si284 as the training set and dev93 as the validation set. For CHiME-4, we use the single channel simulated noisy and the real
Table 1. The CNN architecture for the CNN-BLSTM model

<table>
<thead>
<tr>
<th></th>
<th>in_channel</th>
<th>out_channel</th>
<th>kernel</th>
<th>stride</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv</td>
<td>1</td>
<td>64</td>
<td>$3 \times 3$</td>
<td>1</td>
</tr>
<tr>
<td>conv</td>
<td>64</td>
<td>64</td>
<td>$3 \times 3$</td>
<td>1</td>
</tr>
<tr>
<td>maxpool</td>
<td>64</td>
<td>128</td>
<td>$2 \times 2$</td>
<td>2</td>
</tr>
<tr>
<td>conv</td>
<td>64</td>
<td>128</td>
<td>$3 \times 3$</td>
<td>1</td>
</tr>
<tr>
<td>maxpool</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2 shows the performance of the proposed TL approach to freeze the clean classifier trained in WSJ, reinitialize randomly and retrain the feature extractor using CHiME-4 data. When the top two layers (the softmax layer and the topmost BLSTM layer with its following linear layer) are frozen, the model gives significantly smaller CER compared to the CNN-BLSTM model trained only using CHiME-4. We also notice that if only the softmax layer is frozen, the model does not outperform the baseline, which implies the frozen softmax does not have the capacity to force the feature extractor to learn better features. On the other hand, although freezing three layers surpasses the baseline, it gives inferior results compared to freezing two layers, which indicates that although the three-layered classifier has more capacity, the shallower feature extractor is not powerful/flexible enough to well fit the classifier. The capacity of the classifier and the feature extractor are well balanced when the top two layers are frozen.

5.1. Random reinitialization of the feature extractor

Table 2. Character error rate (CER) of different models. No transfer learning means the models are trained only using CHiME-4. The results of BLSTM CTC is from a previous work [4].

5.2. Pre-training of the feature extractor

Here we consider the case of pre-training the feature extractor using WSJ and further training using CHiME-4. That is, the feature extractor is not randomly reinitialized. Instead, it
<table>
<thead>
<tr>
<th>Model/CER</th>
<th>airport wv1</th>
<th>babble wv1</th>
<th>car wv1</th>
<th>clean wv1</th>
<th>restaurant wv1</th>
<th>street wv1</th>
<th>train wv1</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model A</td>
<td>11.3</td>
<td>12.0</td>
<td>9.8</td>
<td>8.1</td>
<td>13.1</td>
<td>12.7</td>
<td>13.6</td>
<td>11.5</td>
</tr>
<tr>
<td>Model B</td>
<td>12.2</td>
<td>13.1</td>
<td>11.2</td>
<td>9.7</td>
<td>14.1</td>
<td>14.0</td>
<td>14.6</td>
<td>12.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model/CER</th>
<th>airport wv2</th>
<th>babble wv2</th>
<th>car wv2</th>
<th>clean wv2</th>
<th>restaurant wv2</th>
<th>street wv2</th>
<th>train wv2</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model A</td>
<td>26.6</td>
<td>27.3</td>
<td>24.1</td>
<td>22.1</td>
<td>27.4</td>
<td>27.9</td>
<td>28.2</td>
<td>26.2</td>
</tr>
<tr>
<td>Model B</td>
<td>30.5</td>
<td>30.8</td>
<td>26.9</td>
<td>25.2</td>
<td>31.5</td>
<td>31.2</td>
<td>32.5</td>
<td>29.8</td>
</tr>
</tbody>
</table>

### Table 3
Character error rate on all 14 test sets of Aurora-4. Both model A and model B are initialized using the clean CTC model trained on WSJ. For model A, the learning rate for the top two layers (the softmax layer and the topmost BLSTM layer with its following linear layer) are scaled by a factor of 0.5. For model B, there is no learning rate re-scale. wv1 means the utterances are recorded by the primary microphone and wv2 means the utterances are recorded by the secondary microphone.

<table>
<thead>
<tr>
<th>Model/CER</th>
<th>dt05_Multi</th>
<th>et05_real</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier: One layer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frozen</td>
<td>22.0</td>
<td>33.8</td>
</tr>
<tr>
<td>LR scaled by 0.1</td>
<td>22.2</td>
<td>33.5</td>
</tr>
<tr>
<td>LR scaled by 0.5</td>
<td>22.6</td>
<td>34.2</td>
</tr>
<tr>
<td>Classifier: two layers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frozen</td>
<td>22.5</td>
<td>33.8</td>
</tr>
<tr>
<td>LR scaled by 0.1</td>
<td>22.0</td>
<td>33.3</td>
</tr>
<tr>
<td>LR scaled by 0.5 (Model A)</td>
<td><strong>21.9</strong></td>
<td><strong>32.9</strong></td>
</tr>
<tr>
<td>Classifier: three layers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frozen</td>
<td>24.6</td>
<td>36.6</td>
</tr>
<tr>
<td>LR scaled by 0.1</td>
<td>22.7</td>
<td>34.5</td>
</tr>
<tr>
<td>LR scaled by 0.5</td>
<td>21.9</td>
<td>33.4</td>
</tr>
<tr>
<td>No LR re-scale (Model B)</td>
<td>21.9</td>
<td>33.3</td>
</tr>
<tr>
<td>LR scaled by 0.5</td>
<td>22.1</td>
<td>33.4</td>
</tr>
<tr>
<td>for all layers (Model C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No transfer learning</td>
<td>29.0</td>
<td>38.7</td>
</tr>
</tbody>
</table>

### Table 4
Character error rate (CER) for using the clean CTC model to initialize the training on CHiME-4. The classifier is made of the topmost layer(s). During the training on CHiME-4, the classifier is either frozen or the learning rate (LR) is scaled by a small factor. No transfer learning means the model is trained only using CHiME-4.

In this paper, we present a novel transfer learning method from clean data to noisy data for speech recognition. In our proposed method, a clean model is firstly trained on clean data. Then, the clean classifier of the clean model (the top layers) are either frozen or trained with a small learning rate on the noisy data. The feature extractor (the bottom layers) is trained on the noisy data with no learning rate re-scale. Our experiment results on CTC models show our method forces the feature extractor to learn noise invariant features and leads to significant character error rate reductions. Our proposed method is not only constrained to CTC models. Testing our transfer learning methods for other models is left as a further work.
7. REFERENCES


