INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Label Driven Time-Frequency Masking for Robust Continuous Speech Recognition

Meet Soni, Ashish Panda

TCS Innovations Lab
Yantra Park, Thane, Mumbai, India
{meet.soni, ashish.panda}@tcs.com

Abstract

The application of Time-Frequency (T-F) masking based approaches for Automatic Speech Recognition has been shown to provide significant gains in system performance in the presence of additive noise. Such approaches give performance improvement when the T-F masking front-end is trained jointly with the acoustic model. However, such systems still rely on a pre-trained T-F masking enhancement block, trained using pairs of clean and noisy speech signals. Pre-training is necessary due to large number of parameters associated with the enhancement network. In this paper, we propose a flat-start joint training of a network that has both a T-F masking based enhancement block and a phoneme classification block. In particular, we use fully convolutional network as an enhancement front-end to reduce the number of parameters. We train the network by jointly updating the parameters of both these blocks using tied Context-Dependent phoneme states as targets. We observe that pre-training of the proposed enhancement block is not necessary for the convergence. In fact, the proposed flat-start joint training converges faster than the baseline multi-condition trained model. The experiments performed on Aurora-4 database show 7.06% relative improvement over multi-conditioned baseline. We get similar improvements for unseen test conditions as well.

Index Terms: speech recognition, Time-Frequency making, robust speech recognition, multi-conditioned training

1. Introduction

Deep Neural Network (DNN) based systems have become prevalent for Automatic Speech Recognition (ASR). The current ASR systems work well in relatively clean conditions. Now the focus is on improving the robustness of such systems to various degradation conditions such as channel distortion, presence of additive noise, reverberation etc. [1]. Even the best ASR systems trained on clean data performs poorly on noisy data. This is attributed to the mismatch between train and test conditions. It has been shown in literature that DNN-based acoustic models can adapt well if trained using degraded data [2]. Such training is referred to as multi-condition training (MCT).

Many approaches have been proposed to improve the performance of MCT over the years. [3] first proposed the joint training of Time-Frequency (T-F) masking based enhancement front-end and DNN based acoustic model. In this work, a pre-trained T-F masking enhancement network was used along with a DNN-based acoustic model. Both the networks were then combined and their parameters were updated jointly by using Context-Dependent (C-D) phoneme labels as the overall network output. This work was extended in [4] by employing additional features and more in-depth analysis, and in [5] for large vocabulary task by employing Long Short-Term Memory (LSTM) network as both the enhancement front-end and phoneme classification network. Similarly, [6] proposed the joint training of an enhancement network and phoneme classification network by directly predicting the enhanced features without using T-F masking. In [6], the authors used a pre-trained feature enhancement network to estimate the clean features given noisy features. The joint training was similarly performed as [3, 4] using the pre-trained feature enhancement network. In [7], it was shown that, in the context of speech separation, T-F masking with ratio masks is perhaps the most effective way to handle background noise when using DNNs. This was attributed to the the inherently normalized and bounded nature of the ratio masks and also to the fact that denoising via masking is less sensitive to estimation errors.

All these approaches improves upon the MCT baseline by a significant margin. However, this improvement comes at the cost of pre-training a T-F masking based or a direct feature enhancement network. This pre-training requires parallel corpus containing clean-degraded speech utterances. Moreover, such approaches result in larger training times for overall network training. In this paper, we propose a joint training framework based on T-F masking enhancement that does not require pre-trained enhancement network. Therefore, it does not require parallel corpus of clean-degraded speech. Also, since pre-training the enhancement network is not required, proposed approach results in significantly faster training time.

2. Relation to Prior Work

It has been shown that enhanced features that are ideal for ASR may not be ideal for perfectly resynthesizing the target clean speech [5, 8]. And joint training of enhancement network and phoneme classification network is more appropriate strategy to improve the ASR performance [3–6]. However, such jointly trained networks still rely on pre-trained feature enhancement networks. Such pre-training is done with the objective of minimizing the Mean Square Error (MSE) between network output and clean speech features. It has been shown that the features enhanced using such training are not optimal for ASR performance [9, 10], and further tuning of network parameters is necessary to improve ASR performance. In such a scenario, the pre-training of enhancement network using such objective function seems counter-intuitive.

However, pre-training of the enhancement network is necessary due to large number of parameters associated with the network architecture [3–6]. In previous studies different neural network architectures such as DNN [3, 4, 6] and Long Short-Term Memory (LSTM) network [5] have been used as the enhancement front-end. Due to the large number of parameters associated with these architectures, it is necessary to employ pre-training for convergence of network parameters. In this paper, we propose joint training of an acoustic model that has a T-

http://dx.doi.org/10.21437/Interspeech.2019-2172

F masking enhancement network and a phoneme classification network. The novelty of our work lies in using a Fully Convolutional Network (FCN) as the enhancement network. FCN has been recently shown to provide better enhancement performance than DNN and LSTM with significantly less number of parameters [11]. Due to less number of parameters associated with the FCN, we envisage that the acoustic model can now be trained jointly with random initialization and can converge without pre-training. In such training, no parallel clean-noisy speech data is required for training of the enhancement block. The entire network is trained using C-D phoneme labels as the network output with random initialization. We believe that in the absence of clean data available for training, the T-F mask learned by the enhancement network will be optimal for improving ASR performance.

Experiments performed on Aurora-4 database shows that the proposed network converges with random initialization of both enhancement and phoneme classification network. The T-F masking enhancement driven by ASR labels provide significant performance improvement over the MCT baseline. We show that the performance improvement is in fact due to the T-F masking enhancement formulation in the network architecture and not because of simply adding more number of parameters (or layers) to the network.

3. Label Driven Time-Frequency Masking

The network architecture of the proposed acoustic model consists of two blocks. The first block is a T-F masking enhancement front-end, which is a Fully Convolutional Network (FCN). The input to the FCN is the noisy log Mel filterbank Energies (MFBEs). We design the enhancement network in a way that it gives the same dimensional output as the input. To achieve this, we use padded convolution without any pooling operation in convolutional layers. The output of the T-F masking enhancement network is a T-F mask that can be applied to the noisy T-F representation to get the enhanced T-F representation. The mask values are restricted to have values between 0-1 to emulate the effects of Ideal Ratio Mask (IRM) [10].

The second block is a phoneme classification network. In general, any network architecture can be used as a phoneme classification block since the output of the enhancement block is of the same dimension as the input. In this work, we use a fully connected DNN as the phoneme classification model to compare the performances with the established DNN baselines. The T-F mask obtained as the output of the enhancement block is applied to the input T-F representation to form an enhanced T-F representation. This enhanced representation is then given as the input to the phoneme classification block by reshaping it. Mathematically, the operations can be represented as follows,

$$\hat{Y}(t,f) = log\{exp(Y(t,f)) \circ M(t,f)\}, \qquad (1)$$

where $Y(t,f)$ is the log-magnitude of noisy T-F representation, $M(t,f)$ is the T-F mask obtained as the output of the enhancement block, and $\hat{Y}(t,f)$ is the enhanced T-F representation in log-magnitude domain. $\hat{Y}(t,f)$ is taken as the input to the phoneme classification DNN. The output of the DNN is the C-D phones of the HMM. The entire network is trained to optimize the Cross-Entrpy (CE) loss between true C-D phone label and the output of the acoustic model. The parameters associated with both the enhancement block and phoneme classification block were initialized randomly. Hence, the resulting T-F mask will be learned in the manner that will maximize the correct classification probability.

In previous joint training frameworks involving T-F masking [3–5], the enhancement network is trained using parallel clean-degraded speech pairs. In that case it is necessary that the data used for training the enhancement network is different than the data used to train the classification network to ensure better generalization to unseen degradation conditions [5]. The classification network is then trained using either clean speech features or using the output of the enhancement front-end by fixing the parameters of enhancement block. Then in the final stage, parameters of both these block are updated using CE loss. Our proposed acoustic model is trained in a single stage without using any clean data. The lower number of parameters associated with FCN make it possible to train the network with random initialization. This results in considerably less training time and does not require pairs of parallel clean-degraded speech data. As it is noted in [5], the ideal T-F mask for improving the ASR performance need not be similar to the ideal T-F mask for improving the perceptual speech quality. Hence, in the absence of ideal T-F masks the enhancement block will learn the mask that is only optimal for ASR performance. The T-F masking enhancement block also enables the visualization of the learned T-F mask and can give insights in what constitutes an optimal T-F mask for improving the ASR performance.

4. Experiments

4.1. Database description

All the experiments were performed on the Aurora-4 database [12]. Aurora-4 is a medium vocabulary database used for noise robust continuous speech recognition task. It contains speech data in the presence of additive noises and linear convolutional (channel) distortions. It contains two training sets. One is clean training set consisting of 7138 utterances recorded by the primary Sennheiser microphone. The other one is time-synchronized multi-conditioned training set. One half of the utterances were recorded by the primary Sennheiser microphone while the other half were recorded using one of the secondary microphones. Both halves include a combination of clean speech (893 utterances) and speech corrupted by one of six different noises (street, train station, car, babble, restaurant, airport) at 10-20 dB SNR (2676 utterances). Two test sets consists of 330 utterances from 8 speakers, which were recorded by the primary microphone and a set of secondary microphones, respectively. Each set was then corrupted by the same six noises used in the training set at 5-15 dB SNR, creating a total of 14 test sets. These 14 test sets were grouped into 4 subsets: clean (Set 1, denoted by A), noisy (Set 2 to Set 7, denoted by B), clean with channel distortion (Set 8, denoted by C), noisy with channel distortion (Set 9 to Set 14, denoted by D). Moreover, 100 utterances are chosen from development set available with Aurora-4 as a validation set for tuning the parameters associated with the proposed model. These utterances were similarly recorded and corrupted by conditions used in the test set. Hence, giving 1400 total validation utterances.

To evaluate the performance of the proposed method in unseen noise conditions, we create a training set by adding unseen noises (i.e. noise types different from the ones that are found in the test utterances) to the clean utterances of the AURORA-4 database. This ensures that the noise types encountered in the test utterances were unseen during the training phase. We created two multi-conditioned training set by adding 100 types of environmental noises [13], and 11 types of noises from Noisex noise database [14] with the SNR of 0-15 dB with 5 dB of increment, following [15]. We removed the babel noise from Noisex since it is present in the test set. We retain the clean utterances recorded by both primary and secondary microphones with the multi-conditioned data. The resulting training set had 893 clean utterances recorded using primary microphone, 893 clean utterances recorded using secondary microphones, and 5352 utterances with additive noise, recorded using primary microphone.

4.2. Network architecture and training

For ASR system building we use hybrid DNN-HMM acoustic model. First, we develop GMM-HMM system using 13 dimensional Mel-Filter Cepstral Coefficients (MFCCs) in Kaldi [16] using the WSJ recipe. The GMM-HMM system is trained on clean data. Then the alignments of clean data is used to develop DNN-HMM system on multi-conditioned data. DNNs were implemented in Tensorflow. The DNN had 7 hidden layers with 2048 hidden units and ReLU activation. The input layer had 11-frame context of 24 dimensional log-MFBEs. The output layer had 3088 softmax units, corresponding to 3088 tied states of HMMs. Input features were normalized to have zero mean and unit variance. Moreover, the utterance-level mean normalization was also used as suggested in [3]. The network was trained with random initialization for 30 epochs. The batch size of 256 was used and the learning rate was scheduled to decrease linearly as per training epochs. For first 20 epochs, the learning rate was decreased from initial 0.001 to final 0.0001. The final learning rate of 0.0001 was kept constant for remaining epochs. The network was trained using Stochastic Gradient Descent (SGD) with Adam optimizer [17].

The enhancement block was an FCN. We derived the optimal set of parameters such as number of convolutional layers, number of filters in a layer, and filter sizes via parameter search using the validation set. The optimal network had 4 convolutional layers with filter sizes of 5×7 , 5×5, 5×5, and 5×5, respectively. The number of filters in each layer were 60, 60, 60, and 1. Each layer, except for the final layer had ReLU activations. The final layer had sigmoid activation to restrict the mask values between $0 - 1$. Input to the enhancement block was 21-frame context of the log-MFBEs giving input size of 21×24. Since we use padded convolution, the output of the enhancement block was also of the same dimension. However, we use only 11 centre frames as an input to the acoustic model to compare the results with the baseline model. Hence, the 21-frame context is used to obtain the T-F mask, but only 11 frame context is used for ASR. The obtained mask is applied to the input T-F representation as per Eq. 1. This results in the enhanced T-F representation. The center 11 frames were extracted and flattened to use as the input to the DNN. Hence, the input dimensions for DNN were $11 * 24 = 264$. The training parameters for this DNN was same as the baseline model.

We observed that the proposed acoustic model converges faster than the baseline DNN model and quickly starts overfitting. Hence, to choose the best model we perform decoding on the validation set after each epoch and choose the model with the least WER on the validation set. We observed that unlike the baseline DNN model, the proposed model converges to the best performance within 5-6 epochs with same training parameters. Further we verify that the performance improvement is in fact due to T-F masking enhancement network, and not due to additional parameters in the overall network architecture. We do this by training one more network that has similar architecture but without T-F masking formulation shown in Eq. 1. To achieve this we take output of the FCN enhancement block directly as the input to DNN classification network. In this case we treat the output of the FCN enhancement block as the enhanced feature representation and we do not multiply it with the input T-F representation. This network has the similar architecture as the proposed network with one difference. We use linear activation function at the final layer of FCN since the output values need not be restricted between 0-1. All other parameters associated with the network were similar and the training and model selection was also performed similarly.

Table 1: *WER (%) on Aurora-4 test set using Aurora-4 seen multi-condition training set. The results are shown for baseline, proposed network with T-F masking enhancement block and a network with direct feature enhancement block without the T-F masking formulation.*

Noise source	A	B	C	D	Average
Aurora-4 seen MCT	4.01	7.32	9.13	18.65	11.48
Aurora-4 seen Direct	3.97	7.12	6.59	18.61	11.20
Aurora-4 seen T-F mask	3.99	7.41	6.34	17.17	**10.67**

Table 2: *WER (%) on Aurora-4 test set using unseen noises in training set. The results are shown for MCT and proposed network with T-F masking enhancement block. The unseen noises include environmental noises [13], and noises from Noisex [14] databse.*

Noise source	A	B	C	D	Average
Environment MCT	4.02	8.79	9.76	25.4	15.38
Environment T-F mask	3.84	8.37	7.68	24.31	**14.28**
Noisex MCT	4.08	10.19	9.98	24.29	15.34
Noisex T-F mask	4.06	9.81	7.48	24.38	**14.98**

4.3. Results and discussions

Table 1 shows the results of the acoustic models trained on Aurora-4 multi-condition training set. The results are shown for baseline DNN models, acoustic models with the proposed T-F masking enhancement block, and the model with direct feature enhancement, without T-F masking. The baseline model trained on Aurora-4 multi-conditioned data gives 11.48% Word Error Rate (WER). This baseline is stronger than the one reported in [3] for the same feature-set and network architecture. The proposed model gives 10.67% WER with the T-F masking enhancement block. It shows 7.06% relative improvement over the baseline model. While the model with direct enhancement block without T-F masking gives 11.20% WER. It has been shown in [15] that by simply adding more layers to DNN baseline does not improve the ASR performance significantly. The results from the model with direct enhancement block seems to confirm this. Merely adding the FCN enhancment block to the DNN acoustic model did not result in signficant performance gain. However, the proposed model with T-F masking improves the ASR performance by a signficant margin. This result shows the significance of the T-F masking formulation. The highest improvement was observed in C and D test conditions that correspond to channel distortions and channel distortions plus additive noise, respectively. These results are along the similar lines as the results reported in [6] on Aurora-4 database, where performance in C and D test conditions improved significantly with joint training.

Table 2 shows the results of various ASR models trained on different unseen noisy conditions. We got the similar per-

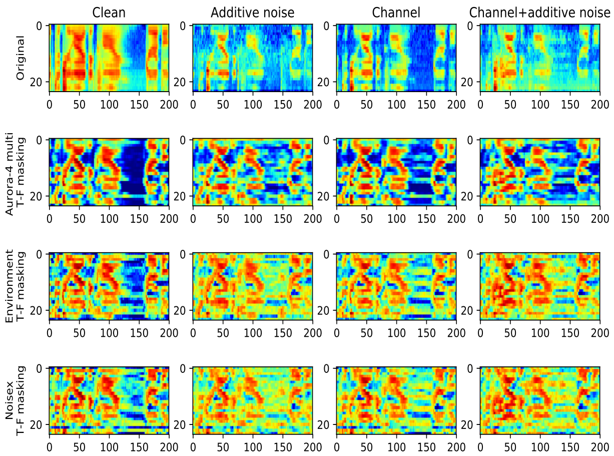

Figure 1: *Noisy as well as enhanced log-MFBEs obtained using T-F masks learned using different training data. The features are shown for a clean utterance and the same utterance with additive noise, channel distortion, and channel plus additive noise. The utterances were taken from Aurora-4 test set.*

formance improvement in the case of unseen noisy conditions. The baseline model trained using environmental noises gave 15.38% WER. While the proposed model using the same data gave 14.28% WER. Similarly, the baseline model trained with noises from NOISEX database gave 15.34% WER. While the proposed model improved the performance to 14.98% WER using the same training data. In the case of environmental noises and noises from NOISEX database, we also got performance improvement in test condition B that constitutes additive noises. The performance did not change significantly in the case of test condition D for this set of experiments. We suspect the reason behind this can be the absence of channel plus additive noise conditions in the training dataset. Moreover, the performance improvement for environmental noises was more than using NOISEX noises. We believe this is due to more number of noises in the training data.

Table 3: *Standard deviation of mean square errors between clean and degraded/enhanced log-MFBEs. Results are shown noisy log-MFBEs and for enhanced features by applying T-F masks obtained using networks jointly trained on different degradation conditions.*

Condition	Original noisy	Aurora-4 multi TF	Environment TF	Noisex TF
Standard deviation	0.304	0.031	0.033	0.035

Next, we observe the properties of T-F masks learned by joint training on various databases. We visualize the enhanced T-F representations after applying T-F masks learned by the enhancement network for various degradation conditions in Figure 1. The first row shows the log-MFBEs of original utterances with various degradation conditions. The second column shows the resultant enhanced log-MFBEs after applying the T-F mask learned by the joint network trained on Aurora-4 multi-condition data. The resultant enhanced T-F representations show that masks generated by the jointly trained model attenuate noise, while preserving spectro-temporal patterns that are most important for recognition. The enhanced T-F representations are different in nature than clean T-F representation, which are expected targets of typical feature enhancement networks. Similar properties can be observed for the enhanced T-F representations after applying T-F masks learned using unseen training noises. In this case, we observe that the T-F masks focuse more on identifying the useful spectro-temporal patterns than attenuating noise. We suspect this behavior is due to mismatch between train and test conditions. However, it should be noted that after T-F masking, the signals under different conditions look more alike than the same signals without T-F masking. This "alikeness" in the signal under different conditions provides the performance improvement seen in the experiments.

We now try to quantify the "alikeness". While comparing the enhanced T-F representations, it can be said that the variance of the enhanced T-F representations across various degradation conditions is significantly less than the original degraded speech. To quantify the variance, we find the mean squared difference between the original clean and degraded features for various degradation conditions and then compute the standard deviation of the difference. For T-F masking based enhanced features, we perform similar calculations after applying the respective T-F masks obtained using jointly trained network. Table 3 shows the resultant standard deviations across conditions. It can be observed that the standard deviation of mean squared error is very high in the case of degraded signals. After enhancing the signals using learned T-F masks, the standard deviation drops significantly. It shows that the learned T-F masks try to normalize the resultant T-F representations for various degradation conditions by suppressing noise and identifying the important T-F points.

5. Conclusions

In this paper we proposed a joint training framework for Multi-Condition Training (MCT) with a T-F masking enhancement block and a phoneme classification block. We show that it is possible to train such network with random initialization, without using any clean data. We achieve this by using Fully Convolutional Network (FCN) as the enhancement block. Due to smaller number of parameters associated with FCN, it is possible to train the T-F mask estimation network jointly with the phoneme classification network with random initialization. We show that such network improves significantly upon MCT baseline without using clean data. Performance improvements we achieve are also consistent over unseen degradation conditions. Moreover, we also show that the improvement is in fact due to T-F masking formulation rather than adding more network parameters. The study of resultant enhanced T-F representations show that the learned T-F masks suppress noise and enhance spectro-temporal patterns responsible for better ASR performance across various degradation conditions.

The proposed framework is generalized and can be applied to any kind of acoustic model. In future, we plan to use different acoustic models like LSTM and CNN along with the proposed T-F masking enhancement network. Moreover, the FCN used in this work consists of vanilla convolutional layers. This can be further improved by using more sophisticated network architecture that includes batch-normalization, skip connections, etc. Varying feature dimensions and context size is also expected to improve the performance using the proposed method.

6. References

[1] V. Mitra, H. Franco, R. M. Stern, J. Van Hout, L. Ferrer, M. Graciarena, W. Wang, D. Vergyri, A. Alwan, and J. H. Hansen, "Robust features in deep-learning-based speech recognition," in *New Era for Robust Speech Recognition*. Springer, 2017, pp. 187–217.

[2] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," *arXiv preprint arXiv:1301.3605*, 2013.

[3] A. Narayanan and D. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2504–2508.

[4] ——, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 23, no. 1, pp. 92–101, 2015.

[5] A. Narayanan, A. Misra, and K. K. Chin, "Large-scale, sequence-discriminative, joint adaptive training for masking-based robust asr," in *INTERSPEECH*, 2015.

[6] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4375–4379.

[7] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[8] A. Narayanan and D. Wang, "The role of binary mask patterns in automatic speech recognition in background noise," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3083–3093, 2013.

[9] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7398–7402.

[10] A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 826–835, 2014.

[11] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," *Proc. Interspeech 2017*, pp. 1993–1997, 2017.

[12] D. Pearce and J. Picone, "Aurora working group: Dsr front end lvcsr evaluation au/384/02," *Inst. for Signal & Inform. Process., Mississippi State Univ., Tech. Rep*, 2002.

[13] G. Hu, "100 nonspeech environmental sounds, 2004."

[14] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[15] Y. Qian, M. Yin, Y. You, and K. Yu, "Multi-task joint-learning of deep neural networks for robust speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 310–316.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.