ENVIRONMENTAL NOISE EMBEDDINGS FOR ROBUST SPEECH RECOGNITION

Suyoun Kim, Bhiksha Raj, Ian Lane

Carnegie Mellon University
Electrical and Computer Engineering
suyoun@cmu.edu, bhiksha@cs.cmu.edu, lane@cmu.edu

ABSTRACT

We propose a novel deep neural network architecture for speech recognition that explicitly employs knowledge of the background environmental noise within a deep neural network acoustic model. A deep neural network is used to predict the acoustic environment in which the system in being used. The discriminative embedding generated at the bottleneck layer of this network is then concatenated with traditional acoustic features as input to a deep neural network acoustic model. Using simulated acoustic environments we show that the proposed approach significantly improves speech recognition accuracy in noisy and highly reverberant environments, outperforming multi-condition training and multi-task learning for this task.

Index Terms— robust speech recognition, deep neural network, noise embeddings

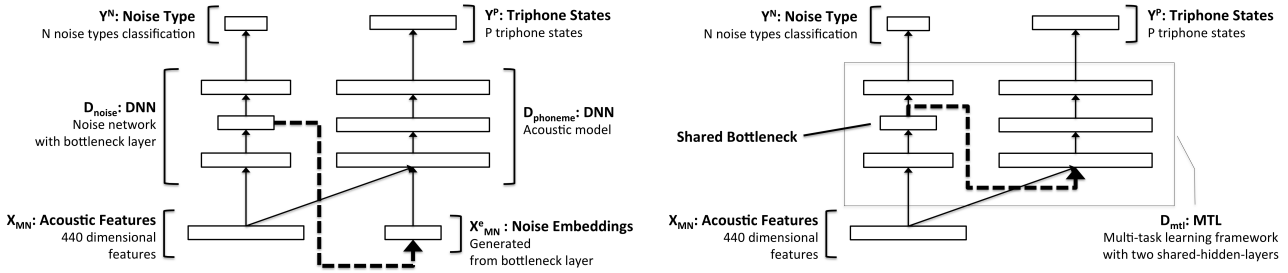
1. INTRODUCTION

In many speech recognition tasks, despite an increase in the variability of the training data, it is still common to have significant mismatches between test environment and training environment, e.g. ambient noise and reverberation. This environmental distortion results in the performance degradation of automatic speech recognition (ASR). Various techniques have been introduced for increasing robustness in this situation.

Over the years, prior works on improving robustness under environmental distortion has generally fallen into three categories: feature enhancement, transformation, and model generalization. Feature enhancement approaches try to attenuate the corrupting noise in the observation and develop more robust feature representation in order to minimize the mismatches between training and test conditions. Many of these methods have been proposed to suppress noise, for example, the model-based compensation methods, Vector Taylor Series (VTS), attempt to model the nonlinear environment function and then apply the compensation for the effects of noise [1], the noise robust feature extraction algorithms based on the different characteristics of speech and background noise have been developed [2, 3], and the missing feature approaches, attempt to mask or impute the unreliable regions of the spectral components because of degradation due to noise have been proposed [4, 5, 6]. Transformation approaches attempt to transform the feature or model space adaptively according to each speaker or each utterance [7, 8]. Generalization approaches attempt to statistically generalize acoustic models by adding noise intentionally or using raw features, allowing the extraction of more invariant features to the influence of the noise. These approaches have been successfully applied through the advent of the Deep Neural Network (DNN) based acoustic model. The DNN-based acoustic model has been shown to be able to capture useful information that might be discarded by traditional preprocessing and to extract the invariant features [9, 10, 11]. A noise-aware training technique which uses the estimated noise at each input frame has been proposed [12] and a dropout technique that randomly drops units from the network during training has been introduced [13] can be categorized into model generalization approach.

The main focus of these previous works has been to either attenuate noise or to utilize noise for the purpose of better generalization. However, we propose to explicitly employ the knowledge of the environmental noise itself (i.e. in which conversations are taking place). Our model incorporates environmental acoustics while training the DNN-based acoustic model in order to improve robustness in environmental distortion. We first build a DNN-based bottleneck model to generate the discriminative acoustic environmental features characterizing specific noise types, which we call *noise embeddings*. We then concatenate these learned embeddings with traditional acoustic features to create input for a DNN acoustic model. Because the noise embeddings are additive information within the network, this is a distinctive approach from the multi-condition training or noise-aware training methods [12]. Through a series of experiments on Resource Management (RM), Wall Street Journal (WSJ0), and Aurora4 datasets [14], we show that our proposed approach improves speech recognition accuracy in various types of noisy environments. In addition, we also compare our approach with the multi-condition training technique [12] and a multi-task learning framework that jointly predicts noise type and context-dependent triphone states.

The paper is organized as follow. In Section 2 we propose



(a) +NE (b) MTL

Fig. 1: Illustration of our approach noise embedding adaptive training +NE and MTL framework. (a)+NE is sequentially training two parts of the same network: (1) train environmental embeddings, then (2) train the triphone network. By contrast, (b)MTL is jointly optimized the two components of the network

our strategies to improve noise robustness. In Section 3, we evaluate the performance of the proposed approach. Finally, we draw conclusions in Section 4.

2. ENVIRONMENTAL NOISE EMBEDDINGS

2.1. Learning environmental noise embeddings

As a first step, we learn the noise embeddings from a narrow bottleneck hidden layer in DNNs, given various types of noisy speech data. A bottleneck neural network is a kind of multi-layer perceptron (MLP) in which one of the internal layers has a small number of hidden units, relative to the size of the other layers [15, 16, 17, 18]. It has been shown that the features generated from the bottleneck network can be classified into a low dimensional representation by forcing this small layer to create a constriction in the network, and consequently it can be represented as a nonlinear transformation and leads to dimensionality reduction of the input features. We take advantage of this fact to generate the low dimensional secondary feature vector.

To make the bottleneck feature vector embed the discriminative acoustic characteristics of background noise, not the speech acoustics, the task of the network is to classify different noise conditions. We start with multiple speech data under the different noise conditions: Acoustic feature, X_{MN}, where M is the number of input frames and N is the number of noise conditions. X_{MN} maps to the two different ground-truth categorical labels: $Y^P_{\{1...i\}}$ and $Y^N_{\{1...i\}}$, where an index of input $i \in \{1, ..., M\}$. $Y^P_{\{1...i\}}$ represents the context-dependent triphone state and $Y^N_{\{1...i\}}$ represents the type of noise. The two separate neural networks, D_{noise} and $D_{phoneme}$, are built from a subset of X_{MN} in order. First, D_{noise}, is optimized with respect to the objective function,

$$F_{noise} = -\sum_{j=1}^{i} logP(Y_j^N | X_{jN}, \theta) \quad (1)$$

After the parameters of the bottleneck network converged, the noise embeddings X^e_{MN} corresponding to each original input X_{MN} are extracted from the network D_{noise}.

2.2. Adaptive DNN training

As a second step, the noise embeddings X^e_{MN} are simply concatenated with the original acoustic features X_{MN}, and the final acoustic model $D_{phoneme}$ is trained with the conventional DNN-based acoustic model procedure to predict context-dependent triphone states, $Y^P_{\{1...i\}}$.

$$X^*_{MN} = X_{MN} \frown X^e_{MN} \quad (2)$$

$$F_{phoneme} = -\sum_{j=1}^{i} logP(Y_j^P | X^*_{jN}, \theta) \quad (3)$$

During decoding, we use the estimated noise embeddings from the bottleneck network without any prior knowledge of noise conditions. The estimated noise embeddings for each input frame play a role in adapting to the different noisy environment. As such, our model can deal with environmental distortion dynamically. The Figure 1a illustrates the overall architecture.

2.3. Multi-task learning

We recognize that our framework described in Section 2 is sequentially training two parts of the same network. First we train the environmental embeddings, and then we fix it and train the triphone network. As a comparator, we also attempt joint optimization. Here the two components of the network are jointly optimized. This joint optimization approach can be effectively a multi-task learning setup which is a method that jointly learns more than one problem together at the same time using shared representation. It has been applied to various speech-related tasks, and our setup MTL is similar to these other multi-task learning solutions [19], except that we are considering environment as the variable.

Figure 1b shows the architecture of our MTL approach. We jointly optimize the network to predict the noise label while to predict the triphone states, so that the network can learn noise-related structure. As a secondary task, the noise label classification task is designed to predict the acoustic environmental type $Y^N_{\{1...i\}}$ from the current acoustic observation X_{MN}. For the fair comparison to our framework, +NE, we build the same size of the network in which the two hidden layers are shared across two different task. Especially we make the second shared-hidden-layer has the same dimension as that of our noise embedding feature, so that this second shared-hidden-layer can serve as environmental noise information. Once the network is optimized to minimize both the noise prediction error and the triphone states error, two shared-hidden-layers and the right side of three hidden layers are used for the decoding.

3. EXPERIMENTS

We investigate the performance of our noise embedding technique on three different databases, RM, WSJ0, and Aurora4 [14], in two main ways: in-domain noise experiment, and unseen experiment. In-domain noise experiment, we perform the experiments on the test set with the same types of noises when the model is trained. On the other hand, in the unseen noise experiment, we evaluate the model with test set with unseen noises.

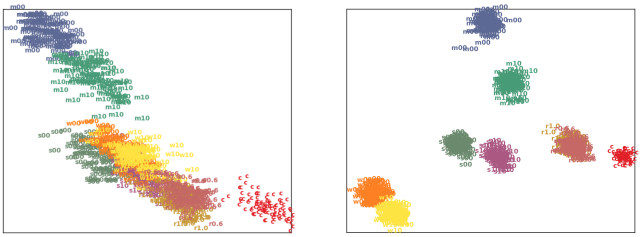
As a first step, we built GMM model by using Kaldi toolkit [20] with their standard recipe, then, we constructing DNN model by using PDNN [21]. We used 11 neighboring frames of 40-dimensional log-scale filterbank coefficients as input, and alignments generated by the GMM model as labels.

For our baseline system, we used a multi-condition training method which enables the network to learn higher level features that are more invariant to the effects of noise [12]. The network contains three hidden layers have 1,064 units for each, it uses 4.86 million parameters. We trained the network using the cross-entropy objective with mini-batch based stochastic gradient descent (SGD) and using the newbob learning rate schedule.

For +NE, we built a DNN that has a narrow bottleneck hidden layer, allowing for the extraction of more tractable, high-level context information. It has three hidden layers. The first and third layer have 500 units for each, while the second layer is a bottleneck with smaller units (i.e. 20 or 50). Once the network was optimized, the discriminative noise embedding features of every training and test sets were concatenated to each corresponding original feature set. Again, these noise embedding features were focused on capturing the background information optimized by different objectives, in contrast with the estimated noise [12], which was focused on generalizing the model to predict context-dependent triphone states by adding noise for the invariant features.

For the multi-task learning system, MTL, we shared two layers as described in Figure 1b. Each layer had 1,024 and 20 hidden units in both tasks. We also had one more hidden layer with 1,024 hidden units for the type of noise classification task, and three more layers with 1,024 hidden units for the triphone prediction task.

For the fair comparison, three models matched approximately same number of parameters. baseline uses 4.86 million parameters, MTL uses 4.87 million parameters, and +NE uses 4.85 million parameters for building the model.



(a) (b)

Fig. 2: Comparison of embeddings of input features between baseline model and +NE model. Left figure (a) shows 100 number of 440-dimensional input features of baseline model and right figure (b) shows 100 number of 450-dimensional input features of +NE model. The colors represent each type of noise condition.

3.1. Simulated in-domain noisy RM datset

We first evaluated our method on the in-domain experiments on the noisy data that have been derived from RM. We artificially mixed the clean speech with eight different types of noisy background, including: white noise at 0 dB, and 10 dB SNR, street noise at 0 dB, and 10 dB SNR, background music at 0 dB, and 10 dB, and simulated reverberation with 1.0 s reverberation time and 600 ms reverberation time. The street noise and the background music segments was obtained from [2], and the reverberation simulations were accomplished using the *Room Impulse Response* open source package [22], and the virtual room size was 5 x 4 x 6 meters.

Figure 2 shows the 450-dimensional final input feature with the learned noise embeddings compared to the 440-dimensional features of the baseline model. The figure shows that adding noise embeddings helps the input feature set be more discriminative with respect to the different environments.

Table 1 compares the recognition accuracy obtained using three models: baseline, MTL, and +NE. It can be seen that at all SNRs and all noise types +NE outperforms the others even in clean datasets. We note that the improvements in recognition accuracy are greater at the lower SNRs. For example, we obtained 2.92 % of WER improvement in the dataset with background music at 0 dB SNR, whereas only 0.19 % of WER improvement in the clean dataset.

Table 1: Comparison of WERs(%) between the baseline system, with NE, and MTL model using 50-dimensional embeddings for 8 different noisy evaluation sets and one clean evaluation set.

Testset(SNR)	baseline	+NE	MLT
clean	3.0	**2.9**	3.1
music(00)	28.4	**25.5**	29.1
music(10)	6.5	**6.3**	7.4
reverb(0.6)	16.4	**15.4**	17.4
reverb(1.0)	26.8	**25.3**	29.0
street(00)	35.0	**32.7**	39.1
street(10)	7.7	**6.7**	7.7
white(00)	30.7	**28.8**	33.8
white(10)	9.7	**8.3**	9.5
Average	18.3	**16.9**	19.5

Table 2: Comparison of WERs(%) on simulated WSJ task and unseen noise using Aurora4 evaluation set between the baseline system, with NE, and MTL model.

Testset(SNR)	baseline	with NE	MTL
WSJ (In-domain Noise)	43.3	41.5	44.6
Aurora4 (Unseen Noise)	37.0	36.7	39.1

3.2. Simulated in-domain noisy WSJ0 datset

We also evaluate our system on the medium size vocabulary WSJ task. Similar to RM experimental setup described in Section 3.1, three additional types of noisy data were generated in addition to clean WSJ0 data: (1) white noise, (2) background music, and (3) street noise. Table 2 compares the WER obtained the baseline system, MTL, and the +NE model. We note that despite the three network sizes were same, +NE provided substantially better recognition error rates than the rates obtained from Baseline and MTL. A 4.34% reduction in relative error rate was obtained by using +NE with 20-dimensional noise embeddings, showing that we can leverage noise embeddings learned with different objective separately to train the acoustic model to adapt to environmental noise type in the medium size task.

3.3. Out-of-domain noise experiments: Unseen noise

In general, the utility of the diverse environment during training is limited by the lack of prior information of noise types, we evaluated our approach in unseen noise conditions. The model trained on WSJ0 corrupted by only three different types of noise (music/street/white) as described in Section 3.1, and then tested with the evaluation set of the Aurora4 dataset [14]. Aurora4 dataset is the 5000-word vocabulary task based on the WSJ0 corpus, and consists 4.84 hours of 2,324 noisy utterances from the Nov'92 Eval Set, including speech corrupted by one of six different noises (street traffic, train station, car, babble, restaurant, airport) at 10-20 dB SNR. The noise embeddings for the evaluation set of Aurora4 were extracted from the network optimized on the corrupted WSJ0 dataset without any environment information of the Aurora4 dataset. Table 2 compares WER obtained using Baseline, MTL, and +NE. The results show that our approach +NE is superior even in the out-of-domain noise situation. Although the improvement of out-of-domain noise case (relative improvement: 0.67%) is less than the gain of in-domain noise case (relative improvement: 4.34%), it is clear that our approach +NE trained with noise embeddings provides better accuracy than Baseline and MTL.

4. CONCLUSIONS

We proposed a noise embeddings adaptive training, +NE to improve robustness under environmental distortion. In the context of the CD-DNN-HMM LVSR framework, we verified the effectiveness of our proposed framework, +NE, by the improved recognition accuracy in all of the noisy conditions, even in clean and unseen noisy conditions. We also demonstrated that the sequentially learned noise embeddings is more effective than the simultaneously learned noise embeddings within the multi-task learning framework.

The implication of this work is significant and far-reaching. First, it suggests the possibility to build a highly robust DNN-based acoustic model in various unseen noisy environment from few known noise environment information. This huge benefit would require a small set of noise types, although having more various noisy data would further improve the performance. In our current study, we only used three noise types to extract noise embeddings. We believe further performance improvement can be achieved by using additional and more diversified noises to cover a wider range of the noise variations. Second, our result in comparison with MTL indicates that our approach, the sequential optimization, can be more effective than the joint optimization in both in-domain and out-of-domain task. This finding is applicable to tasks that need to be optimized for multiple objectives.

In this study we showed that the background environmental noise information trained with separate objectives, we can obtain additional gain by further adjusting the full DNN. We believe this is an indication that various context information should be considered to model the greater robustness in various real world noisy situations.

5. ACKNOWLEDGMENT

The authors would like to acknowledge the contributions made by Richard M. Stern for his valuable and constructive suggestions during the planning and development of this project.

6. REFERENCES

[1] Pedro J Moreno, Bhiksha Raj, and Richard M Stern, "A vector taylor series approach for environment-independent speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. IEEE, 1996, vol. 2, pp. 733–736.

[2] Chanwoo Kim and Richard M Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4101–4104.

[3] Chanwoo Kim and Richard M Stern, "Nonlinear enhancement of onset for robust speech recognition.," in *INTERSPEECH*, 2010, pp. 2058–2061.

[4] Bhiksha Raj and Richard M Stern, "Missing-feature approaches in speech recognition," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 101–116, 2005.

[5] Bo Li and Khe Chai Sim, "Improving robustness of deep neural networks via spectral masking for automatic speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 279–284.

[6] Arun Narayanan and DeLiang 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.

[7] Mark JF Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[8] Mark JF Gales, "Semi-tied covariance matrices for hidden markov models," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, no. 3, pp. 272–281, 1999.

[9] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton, "Acoustic modeling using deep belief networks," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 14–22, 2012.

[10] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[11] Frank Seide, Gang Li, and Dong Yu, "Conversational speech transcription using context-dependent deep neural networks.," in *Interspeech*, 2011, pp. 437–440.

[12] Michael L Seltzer, Dong Yu, and Yongqiang 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.

[13] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[14] N Parihar and J Picone, "Aurora working group: Dsr front end lvcsr evaluation au/384/02," *Inst. for Signal and Information Process, Mississippi State University, Tech. Rep*, vol. 40, pp. 94, 2002.

[15] Dong Yu and Michael L Seltzer, "Improved bottleneck features using pretrained deep neural networks.," in *INTERSPEECH*, 2011, vol. 237, p. 240.

[16] Sibel Yaman, Jason Pelecanos, and Ruhi Sarikaya, "Bottleneck features for speaker recognition," in *Odyssey*, 2012, vol. 12, pp. 105–108.

[17] Jonas Gehring, Yajie Miao, Florian Metze, and Alex Waibel, "Extracting deep bottleneck features using stacked auto-encoders," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3377–3381.

[18] Jonas Gehring, Wonkyum Lee, Kevin Kilgour, Ian R Lane, Yajie Miao, Alex Waibel, and Silicon Valley Campus, "Modular combination of deep neural networks for acoustic modeling.," in *INTERSPEECH*, 2013, pp. 94–98.

[19] Michael L Seltzer and Jasha Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6965–6969.

[20] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.

[21] Yajie Miao, "Kaldi+ pdnn: building dnn-based asr systems with kaldi and pdnn," *arXiv preprint arXiv:1401.6984*, 2014.

[22] Emanuel AP Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, pp. 1, 2006.