CROSS-ACOUSTIC TRANSFER LEARNING FOR SOUND EVENT CLASSIFICATION

Hyungjun Lim, Myung Jong Kim, Hoirin Kim

School of Electrical Engineering, KAIST, Daejeon 34141, South Korea
E-mail: {hyungjun.lim, myungjong, hoirkim}@kaist.ac.kr

ABSTRACT

A well-trained acoustic model that effectively captures the characteristics of sound events is a critical factor to develop more reliable system for sound event classification. Deep neural network (DNN) which has an ability to extract discriminative representation of features can be a good candidate for acoustic model of sound events. Compared to other data such as speech or image, the amount of sound database is often insufficient for learning the DNN properly, resulting in overfitting problems. In this paper, we propose a *cross-acoustic transfer learning* framework that can effectively train the DNN even with insufficient sound data by employing rich speech data. Three datasets are used to evaluate our proposed method; one sound dataset is from Real World Computing Partnership (RWCP) DB and two speech datasets are from Resource Management (RM) and Wall Street Journal (WSJ) DBs. A series of experimental results verify that cross-acoustic transfer learning performs significantly better than the baseline DNN which was trained only from sound data, achieving 26.24% relative classification error rate (CER) improvement over the DNN baseline system.

Index Terms— Sound event classification, speech-to-sound, transfer learning, deep neural network.

1. INTRODUCTION

Recent advances in deep learning have provided remarkable improvements in various recognition/classification applications such as automatic speech recognition (ASR) [1]-[4] and speaker recognition [5]-[7]. In the deep learning framework, a large amount of training data is generally required to learn reliable feature representations. In some situations where only a small amount of data is available, however, we cannot utilize the deep learning framework because it is likely to be overfitted. In such cases, transfer learning [8], [9], which takes advantage of resource-rich data from other domain, can be a good solution.

Researches on transfer learning are mostly conducted in the field of cross-lingual ASR [10]-[14]. Numerous techniques have been applied for cross-lingual acoustic modeling such as Kullback-Liebler divergence based hidden Markov model which is able to exploit multilingual information [10] and subspace Gaussian mixture model with a shared multilingual phonetic subspace [11]. Recently, a deep neural network (DNN) has received great attention in transfer learning due to more abstract and invariant features represented by multiple hidden layers. Das and Hasegawa-Johnson [12] focused on knowledge transfer of a DNN from English to Turkish. Swietojanski *et al*. [13] constructed the acoustic model for German using the pre-trained DNNs from one or all of German, Portuguese, Spanish and Swedish which belong to the same continent, Europe. Huang *et al*. [14] utilized a multilingual DNN concept as an initialization which is jointly pre-trained from several languages such as French, German, Spanish, and Italian. Even though their specific scenarios are different, the main process consists of two parts in common; unsupervised pre-training with resource-rich data followed by supervised fine-tuning with resource-scarce data. Since the pre-training process does not

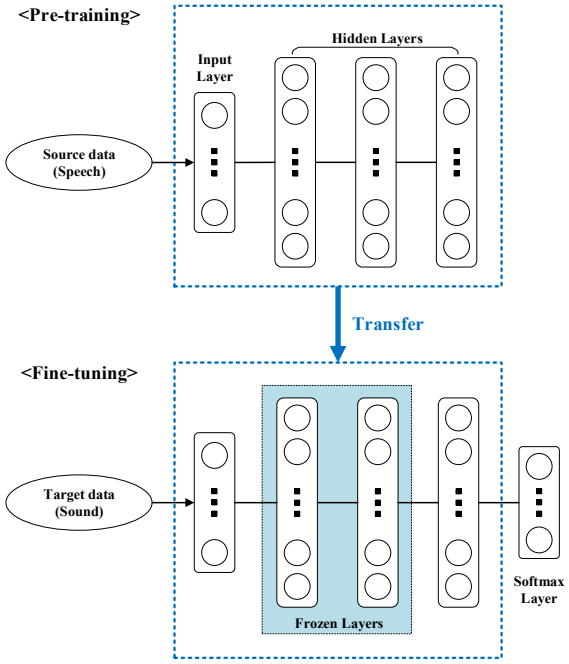

Figure 1. Overall scheme of cross-acoustic transfer learning.

require labeled data, it is possible to use similar data from other domain for better initialization of the DNN.

There are several efforts to apply the deep learning approaches into a sound event classification task including standard DNN [15], composite deep belief network (DBN) [16], and convolutional neural network (CNN) with multiple resolution spectrogram inputs [17]. However, most of them consist of only a small number of hidden layers or hidden units since sound events generally have a relatively short duration compared to other acoustic data such as speech and music. Moreover, in some applications such as audio surveillance, it is hard to collect many data since those sound events (e.g., car crash or explosion) are seldom reproducible in real world. To overcome this problem, we propose a new learning scenario, cross-acoustic transfer learning for sound event classification as depicted in Figure 1. Unlabeled large-scale speech corpus is used to initialize a DNN by unsupervised greedy layer-wise pre-training, and then the pre-trained DNN is further optimized by using supervised fine-tuning with sound data. Finally, our model is compared to baseline DNNs which are trained only with sound data by measuring the classification accuracies of the test sound events. To the best of our knowledge, this work is the first attempt at applying transfer learning to sound event classification using a large amount of speech data.

The remainder of the paper is organized as follows: we provide a brief overview of transfer learning as well as our proposed method in Section 2. In Section 3, several experimental results are presented and our conclusions are summarized in Section 4.

2. TRANSFER LEARNING

Transfer learning is a machine learning technique that transfers knowledge learned from a source domain to a target domain [9]. Even though the 'knowledge' can be interpreted in many different ways, we limit its meaning to 'model parameters' as used in the ASR areas [10]-[14]. In other words, trained model parameters from speech are re-used during the training process for an acoustic model of sound events in this work.

2.1. DNN transfer learning

A DNN is a multi-layer perceptron (MLP) with more than two hidden layers [18], [19]. For sequential data such as speech and sound, an input layer is usually composed of several frames of observations to cover a long context while an output layer consists of several nodes which are identical to the number of labels. Those layers are connected by nonlinearity functions, including sigmoid or hyperbolic tangent which allow a DNN to extract the more distinctive features of input data.

The usual training process of a DNN consists of two parts; an unsupervised generative pre-training which can be done in greedy, layer-wise fashion and a supervised discriminative fine-tuning with error back-propagation. The pre-training process can leverage the DNN to get a better initial point than random initialization for the fine-tuning process by employing a large amount of unlabeled data. Transfer learning which utilize data from different domains can be easily adopted to the DNN training procedure since the pre-training process does not require label information. Therefor a DNN is a suitable framework to apply transfer learning.

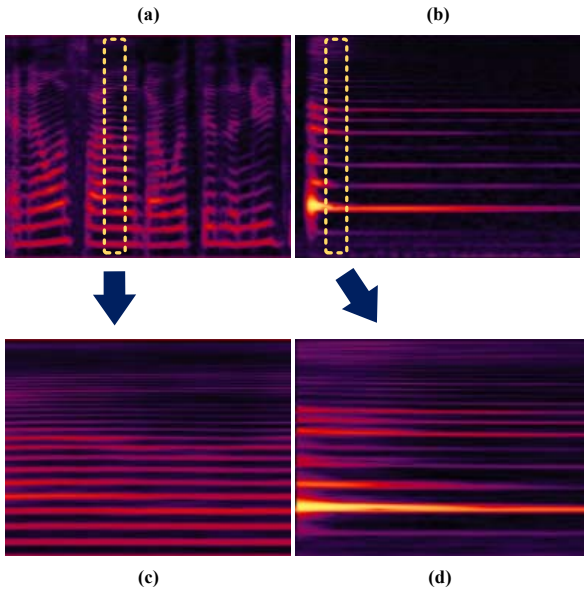

Figure 2. Typical examples of spectrograms; (a) speech (English read speech), (b) sound (twanging of a stringed music instrument), (c) and (d) corresponding enlarged views of the 100 msec context window range, respectively.

2.2. Cross-acoustic transfer learning

Training a DNN for the acoustic model of sound events in the transfer learning framework requires a large amount of database. However, the amount of sound event database usually deficient due to the characteristics of the sound events. A speech database that is generally used in ASR tasks can be a good candidate for this situation since it is sufficient to train a deep model and is similar to sound events in the aspect of acoustic characteristics.

Figure 2 shows typical spectrograms of speech and sound with corresponding enlarged views in the range of 100 msec which is accord with the usual input duration of the DNN for time-domain signals. From Figure 2(c) and (d), we can observe that speech and sound show similar acoustic patterns, i.e., harmonically arranged horizontal lines, even though their entire signals are different. Furthermore, we can analyze that the similarity between speech and sound comes from manner

of articulation of speech [20] and production process of sound like as:

- *Plosive* speech sounds are produced by building up pressure behind a total constriction somewhere in the vocal tract, and suddenly releasing the pressure, which are similar to *explosion* and *collision* sounds.
- *Fricative* speech sounds are produced by exciting the vocal tract by a steady air flow, which are related to *friction between solids*, *air flow*, and *whoosh*.
- *Vowel* speech sounds are produced by exciting a fixed vocal tract with quasi-periodic air pulses, which are similar to *vibration of string or air*.

Therefore, we can utilize speech for training the DNN acoustic model in the transfer learning scenario which is called cross-acoustic transfer learning; the pre-trained DNN with speech data is fine-tuned with sound data.

3. EXPERIMENTS

3.1. Database

In order to evaluate the effectiveness of our method, we selected a total of 50 sound event classes from the Real World Computing Partnership (RWCP) Sound Database in Real Acoustic Environments [21]. The sound events were generated from the wide range of materials and interactions, including *collision of wood*, *metal*, and *plastic*, *dropping particles*, *jetting gas*, *rubbing papers*, and *playing the musical instruments*. Each class contains 100 clips of sounds and we assigned 70 clips for training and 30 clips for testing. Note that each clip is composed of a corresponding sound event as well as short silence before and after the sound event. All sound clips were digitized at 16 bits per sample with 16 kHz sampling rate in the mono-channel. An average duration of each clip is about 1 sec long and totally about 1 hour.

In the proposed transfer learning framework, we used two auxiliary speech corpora for pre-training the DNN; DARPA Resource Management (RM) [22] and Wall Street Journal (WSJ) [23]. We took only training data of those corpora; the RM SI-training set and WSJ SI-284 training set. Finally, we summarize all the databases used in this paper in Table 1.

Table 1. Configuration of three databases**.**

Database	# clips	Total durations
RWCP	5,000	1 hour
RM (SI-training)	4,000	4.4 hours
WSJ (SI-284)	38,000	66 hours

3.2. Baseline DNN

First, we evaluated the performance of the DNN trained from only the RWCP database in the Kaldi+PDNN framework [24], [25]. We used 13 successive 40-dimensional Mel-filterbank log energy as an input of the DNN, which was calculated from 10 msec frame size with 50% overlap. For initialization, a deep belief network (DBN) with Gaussian-Bernoulli restricted Boltzmann machine (GBRBM) for first layer and others are Bernoulli-Bernoulli RBM (BBRBM) was pre-trained with the RWCP training data. After pre-training, a softmax layer was added on top of the DBN, and then fine-tuned with the same data with labels. Note that all tuning parameters including learning rate, number of training epochs, and size of mini-batch were same as default setting in [22]. The classification results of the trained DNN were measured with various number of hidden layers and hidden nodes as shown in Table 2. Here, the best performance is achieved from 2,048 hidden nodes with single hidden layer. As the number of hidden nodes increased from 256 to 2048, we can obtain better results in almost cases. However, the use of multiple hidden layers was not effective in the performance. These results imply that shallow neural network works better than deep architecture when using a small size database as we expected.

Table 2. Classification error rate (%) of baseline DNN system with various hidden layers and nodes (Bold face represents the best result along column axis).

# layers \ # nodes	256	512	1,024	2,048
1	6.75	5.35	4.95	**3.81**
2	6.75	**4.75**	4.88	4.75
3	6.55	5.15	**4.01**	5.28
4	7.62	5.41	5.28	5.08
5	**6.48**	6.02	5.55	4.88
6	8.16	6.28	5.88	5.68

3.3. Transfer learning

In this section, we investigated the effectiveness of cross-acoustic transfer learning by using two different sizes of auxiliary speech datasets, RM and WSJ. To obtain the acoustic model of sound events, unsupervised pre-training with speech data and supervised fine-tuning with sound data were consecutively performed as mentioned in Section 2.

In the usual fine-tuning process of cross-lingual transfer learning, a randomly initialized softmax layer is added on top of the pre-trained DNN, and then just fine-tune the soft-max layer while other layer to be left *frozen* [26]. This is because if we fine-tune a whole DNN, we may fall into the overfitting problem since the size of target database is typically small in transfer learning. Figure 3 shows the classification error rate (CER) of cross-acoustic transfer learning with 6 hidden layers of DNN when the number of frozen layers were varied from 0 to 6. Note that *# of frozen layers* = 0 is the case where the entire hidden layers were fine-tuned as in the standard DNN fine-tuning process. As can be seen, CERs decreased for all cases when we train the more hidden layers during the fine-tuning process both RM and WSJ cases. It means that training more hidden layers is more effective than only softmax layer

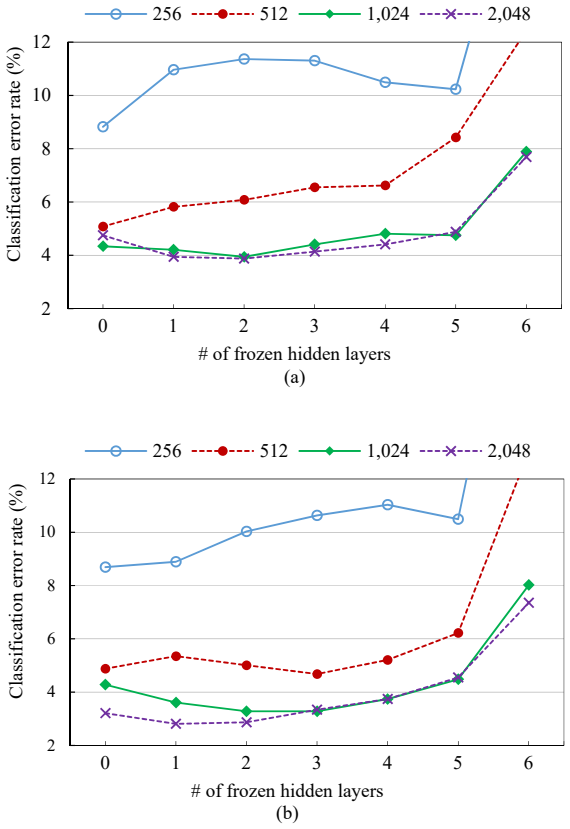

Figure 3. Classification error rate (%) for cross-acoustic transfer learning from two different speech databases with various number of frozen hidden layers: (a) RM, (b) WSJ.

even though the size of target database is small. The trend of the CER performance is opposite to other related works in transfer learning [14], [26]. We can find the reason of those observations from the similarity between source and target data. Since speech and sound are less similar than speech in different languages from the cross-lingual scenario, the initialized DNN with speech data should be more adjusted with sound data for a better representation of sound events. As a result, we can achieve the lowest CER by freezing only a small number of hidden layers.

Next, we compared the CER of cross-acoustic transfer learning from two different speech databases with the baseline DNN listed in Table 3. The results of cross-acoustic transfer learning were coming from the previous experiments which has the lowest CER in each line (See Figure 3). In most cases, cross-acoustic transfer learning outperformed the baseline DNN except 256 hidden nodes case. We can see that the degree of improvement increased as the size of the DNN became larger both RM and WSJ cases. Moreover, we obtained better performance when we used bigger size of pre-

Table 3. Classification error rate (%) for cross-acoustic transfer learning from two different speech databases and the corresponding error reduction rate (%, in parenthesis) compared to the baseline DNN with 6 hidden layers.

# of hidden nodes	256	512	1,024	2,048
Baseline DNN	8.16	6.28	5.88	5.68
Transferred from RM	8.82	5.05	3.94	3.88
(ERR)	(-8.01)	(19.11)	(32.99)	(31.69)
Transferred from WSJ	8.89	4.68	3.28	**2.81**
(ERR)	(-6.50)	(25.48)	(44.22)	(50.53)

training speech database. These results imply that the DNN pre-trained with speech data is very effective to initialize the DNN for sound data. Also, we obtained the best performance of 2.81% CER on WSJ case, achieving 26.24% of relative improvement compared to best performance of the baseline DNN. Finally, we can train a deeper model for resource-scarce sound from our cross-acoustic transfer learning approach which can be able to extract more distinctive feature from sound.

4. CONCLUSIONS

In this paper, we proposed a new learning framework called cross-acoustic transfer learning to train the DNN for modeling resource-scare sound events. Transfer learning was applied to utilize the similarity between speech and sound events for training the DNN. In the method, the DNN was firstly pre-trained with two speech datasets, RM and WSJ, and then fine-tuned with the RWCP sound dataset. In the context of the sound event classification scenario, we conducted a series of experiments to verify the effectiveness of the proposed cross-acoustic transfer learning. The experimental results show that our method outperformed the baseline DNN system, obtaining over 20% relative improvement in terms of CER. Our works would potentially be applied to other resource-scarce environment such as audio surveillance applications. In the future, we will extend our works to train the DNN for modeling the sound events by using additional speech data in various languages and environments, which will cover the wider range of sound events.

5. ACKNOWLEDGEMENT

This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2014R1A2A2A01007650).

6. REFERENCES

[1] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82-97, Nov. 2012.

[2] G.E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Trans*. *Audio*, *Speech*, *and Language Process*., vol. 20, no. 1, pp. 33-42, Jan. 2012.

[3] C. Weng, D. Yu, S. Watanabe, and B.-H.F. Juang, "Recurrent deep neural networks for robust speech recognition," in *Proc*. *IEEE Int. Conf. Acoust. Speech, and Signal Process.,* pp. 5532-5536, May 2014.

[4] Y. Miao and F. Metze, "Distance-aware DNNs for robust speech recognition," in *Proc*. *Interspeech*, pp. 761-765, Sep. 2015.

[5] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc*. *IEEE Int. Conf. Acoust. Speech, and Signal Process.,* pp. 1695-1699, May 2014.

[6] F. Richardson, D.A. Reynolds, and N. Dehak, "A unified deep neural network for speaker and language recognition," in *Proc*. *Interspeech*, pp. 1146-1150, Sep. 2015.

[7] D.G. Romero and A. McCree, "Insight into deep neural networks for speaker recognition," in *Proc*. *Interspeech*, pp. 1141-1145, Sep. 2015.

[8] S.J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans*. *Knowl. Data Eng*., vol. 22, no. 5, pp. 1345-1359, Oct. 2010.

[9] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Trans*. *Audio*, *Speech*, *Language Process*., vol. 21, no. 5, pp.1060-1089, May 2013.

[10] D. Imseng, H. Bourlard, and P.N. Garner, "Using KL-divergence and multilingual information to improve ASR for under-resourced languages," in *Proc*. *IEEE Int. Conf. Acoust. Speech, and Signal Process.,* pp. 4869-4872, May 2012.

[11] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, D. Povey, A. Rastrow, R.C. Rose, and S. Thomas, "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models," in *Proc*. *IEEE Int. Conf. Acoust. Speech, and Signal Process.,* pp. 4334-4337, Mar. 2010.

[12] A. Das and M. Hasegawa-Johnson, "Cross-lingual transfer learning during supervised training in low resource scenarios," in *Proc*. *Interspeech*, pp. 3531-3535, Sep. 2015.

[13] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proc*. *IEEE Workshop* on *Spoken Language Tech.*, pp. 246-251, Nov. 2012.

[14] J.T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc*. *IEEE Int. Conf. Acoust. Speech, and Signal Process.,* pp. 7304-7308, May 2013.

[15] O. Gencoglu, T. Virtanen, and H. Huttunen, "Recognition of acoustic events using deep neural networks," in *Proc*. *IEEE European Signal Process. Conf.*, pp. 506-510, Sep. 2014.

[16] S. Chu, S. Narayanan, and C.C.J. Kuo, "Composite-DBN for recognition of environmental contexts," in *Proc*. *IEEE Asia-Pacific Signal and Inform. Process. Assoc.*, pp. 1-4, Dec. 2012.

[17] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, "Feature extraction strategies in deep learning based acoustic event detection," in *Proc*. *Interspeech*, pp. 2922-2926, Sep. 2015.

[18] L. Deng and D. Yu, *Deep learning*: *Methods and applications*, Now Publishers, 2014.

[19] D. Yu and L. Deng, *Automatic speech recognition – a deep learning approach*, Springer, 2014.

[20] L.R. Rabiner and R.W. Schafer, *Digital processing of speech signals*, Prentice-Hall, 1978.

[21] S. Nakamura, K. Hiyane, F. Asano, T. Yamada, and T. Endo, "Data collection in real acoustical environments for sound scene understanding and hands-free speech recognition," in *Proc*. *Eurospeech*, pp. 2255–2258, 1999.

[22] P. Price, W.M. Fisher, J. Bernstein, and D.S. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," in *Proc*. *IEEE Int. Conf. Acoust. Speech, and Signal Process.,* pp.651-654, 1988.

[23] D.B. Paul and J.M. Baker, "The design for the wall street journal-based CSR corpus," in *Proc*. *Workshop on Speech and Natural Language*, pp. 357-362, 1992.

[24] Y. Miao, "Kaldi+PDNN: building DNN-based ASR systems with Kaldi and PDNN," arXiv:1401.6984, 2014.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc*. *IEEE Autom. Speech Recogn. Understand.,* Dec. 2011.

[26] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc*. *Neural Inform. Process. Syst.*, pp. 3320-3328, Dec. 2014.