UNSUPERVISED DOMAIN ADAPTATION FOR DEEP NEURAL NETWORK BASED VOICE ACTIVITY DETECTION

Xiao-Lei Zhang

Tsinghua National Laboratory for Information Science and Technology,
Department of Electronic Engineering, Tsinghua University, Beijing, China.
huoshan6@126.com

ABSTRACT

The mismatching problem between the training and test speech corpora hinders the practical use of the machine-learning-based voice activity detection (VAD). In this paper, we try to address this problem by the unsupervised domain adaptation techniques, which try to find a shared feature subspace between the mismatching corpora. The denoising deep neural network is used as the learning machine. Three domain adaptation techniques are used for analysis. Experimental results show that the unsupervised domain adaptation technique is promising to the mismatching problem of VAD.

Index Terms— deep learning, domain adaptation, feature learning, transfer learning, voice activity detection.

1. INTRODUCTION

Voice activity detectors (VADs) aim to distinguish noisy speech from the pure noise signals. They are important frontends of modern speech recognition systems [1–4] and speech signal processing systems. Recently, the machine-learning-based VAD [5–14] becomes a promising approach in that it not only can be integrated to the speech recognition systems naturally but also can fuse the advantages of multiple features much better than traditional VADs. However, the machine-learning-based VAD is still far from its practical use. One significant problem is that we are not sure about whether the VAD model trained in one noise scenario is still powerful in a different test noise scenario.

In this paper, we deal with the aforementioned problem by a novel learning method – unsupervised domain adaptation. Generally, unsupervised domain adaptation trains a model with one or multiple labeled source corpora and one unlabeled target corpus, and tests the model on a corpus that is generated from the same underlying distribution as the unlabeled target corpus. See [15] for an excellent survey of the unsupervised domain adaptation. This method is a promising way towards the practical use of the machine-learning-based VAD in that the labeled source corpora are rare and manually expensive, but the model is easily retrained with the extensive power of modern parallel computing systems whenever we encounter new types of noises.

The key idea of our domain adaptation techniques is to extract a low-dimensional feature representation that is shared by the source corpora and the target corpus from multiple acoustic features, by the denoising deep neural networks (DDNN) [14]. Our objective is that the DDNN can generalize well on the test set that has the same noise type as the target corpus in the training set. The main contributions of this paper are summarized as follows:

1. Towards the mismatching problem of the machine-learning-based VAD. Empirical results show that (i) when DDNN is used as the learning machine, the performance of the proposed methods is better than that without the proposed methods. (ii) For a broad comparison, when the distributions of the source noise and target noise are somewhat similar, the proposed methods are more powerful than several referenced VADs. (iii) When the distribution of the source noise is quite dissimilar with the distribution of the target noise, we failed to achieve a good generalization performance on the test set.

2. A useful empirical comparison of three unsupervised domain adaptation schemes. We have proposed three domain adaptations. Empirical results suggest that we would better pretrain the deep neural networks in an unsupervised manner with both the source corpus and the target corpus together, and the data for all layers' pretraining should be consistent without interference.

The remainder of the paper is organized as follows. In Section 2, we present three unsupervised domain adaptation schemes. In Section 3, we present the related work. In Section 4, we conduct an extensive experimental comparison. In Section 5, we conclude this paper with some future work.

2. DOMAIN ADAPTATION FOR VAD

Suppose the training set consists of a labeled source cor-

This work was supported by the China Postdoctoral Science Foundation funded project under Grant 2012M520278.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

Scheme 1 .

```
Pretrain all layers of DDNN with only the unlabeled target corpus $\mathcal{X}^{(t)}$ layer-wisely.
Fine-tune the pretrained DDNN with only the labeled source corpus, i.e. $\mathcal{X}^{(s)} \times \mathcal{Y}^{(s)}$.
```

Scheme 2 .

```
Take the labeled source corpus $\mathcal{X}^{(s)}$ and the unlabeled target corpus $\mathcal{X}^{(t)}$ together as a large corpus, and pretrain all layers of DDNN layer-wisely with the large corpus.
Fine-tune the pretrained DDNN with only $\mathcal{X}^{(s)} \times \mathcal{Y}^{(s)}$.
```

pus $\mathcal{X}^{(s)} \times \mathcal{Y}^{(s)}$ and an unlabeled target corpus $\mathcal{X}^{(t)}$, where $\mathcal{X}$ represents the acoustic feature corpus and $\mathcal{Y}$ represents the set of the manual labels. The corpora $\mathcal{X}^{(s)}$ and $\mathcal{X}^{(t)}$ are sampled from different noise scenarios. Domain adaptation scheme aims to find a mapping function $\phi(\cdot)$ that minimizes the difference between $\phi\left(\mathcal{X}^{(s)}\right)$ and $\phi\left(\mathcal{X}^{(t)}\right)$.

DDNN [14] is used as the learning machine. The key idea of DDNN is to first minimize the *reconstruction cross-entropy loss* between the noisy speech signal and its corresponding clean speech signal in an unsupervised greedy layer-wise pretraining way, and then fine-tune the entire deep neural network by minimizing the *classification cross-entropy loss* between the noisy speech signal and its manual labels.

The core idea of the unsupervised domain adaptation is to first pretrain DDNN in different unsupervised ways and fine-tune DDNN with the labeled source corpus, i.e. $\mathcal{X}^{(s)} \times \mathcal{Y}^{(s)}$. In this paper, we present three unsupervised pretraining schemes, which are listed in Schemes 1, 2, and 3 respectively.

Scheme 1 only uses the unlabeled target corpus $\mathcal{X}^{(t)}$ to pretrain DDNN. It is supposed to be computationally efficient when $\mathcal{X}^{(t)}$ is not very large.

Scheme 2 uses both $\mathcal{X}^{(s)}$ and $\mathcal{X}^{(t)}$ for pretraining DDNN, which can learn a good feature representation shared by $\mathcal{X}^{(s)}$ and $\mathcal{X}^{(t)}$. Particularly, when $\mathcal{X}^{(t)}$ is rare, $\mathcal{X}^{(s)}$ can play a sufficient supplementary role to $\mathcal{X}^{(t)}$. Hence, the network is desired to perform gently well on the test set.

Scheme 3 is designed as a compromise between Scheme 1 and Scheme 2. Specifically, because we inject the supplementary effect of $\mathcal{X}^{(s)}$ merely into the highest layer of DDNN, we might not only transfer the source knowledge to the target domain but also save a lot of training time. Scheme 3 contains two sub-schemes, which is denoted as Scheme $3^{(t)}$ and Scheme $3^{(s)}$ respectively. Scheme $3^{(t)}$ is a scheme that the lowest $L-1$ layers are pretrained by $\mathcal{X}^{(t)}$ only, while Scheme $3^{(s)}$ is a scheme that the lowest $L-1$ layers are pretrained by $\mathcal{X}^{(s)}$ only.

Note that we can use multiple source corpora and multiple target corpora together to train the model freely. But in this paper, we only discuss the empirical performance with one source corpus and one target corpus, leaving the multiple source domain adaptation problem as a future work.

Scheme 3 .

```
Input: The desired depth of DDNN, denoted as $L$, (i.e. the number of the hidden layers).
Source DDNN pretraining: Pretrain a DDNN model with a depth of $L-1$ using only $\mathcal{X}^{(s)}$ as the input. The pretrained source-DDNN is denoted as $\left\{\mathbf{W}_l^{(s)}\right\}_{l=1}^{L-1}$. /*Note: this model needs to be trained only once, and used repeatedly for different target corpus.*/
Target DDNN pretraining: Pretrain another DDNN with a depth of $L-1$ using only $\mathcal{X}^{(t)}$ as the input. The pretrained target DDNN is denoted as $\left\{\mathbf{W}_l^{(t)}\right\}_{l=1}^{L-1}$.
Hybrid pretraining of the top layer: Group the output features of the two DDNN models together to a large training set, and pretrain the $L$-th layer of DDNN with the large set. The pretrained model is denoted as $\mathbf{W}_L^{(t)}$.
/* The following are two model choices.*/
if Scheme $3^{(t)}$ then
    Fine-tune the pretrained model $\left\{\left\{\mathbf{W}_l^{(t)}\right\}_{l=1}^{L-1}, \mathbf{W}_L^{(t)}\right\}$ with only $\mathcal{X}^{(s)} \times \mathcal{Y}^{(s)}$.
else if Scheme $3^{(s)}$ then
    Fine-tune the pretrained model $\left\{\left\{\mathbf{W}_l^{(s)}\right\}_{l=1}^{L-1}, \mathbf{W}_L^{(t)}\right\}$ with only $\mathcal{X}^{(s)} \times \mathcal{Y}^{(s)}$.
end if
```

3. RELATED WORK

The distribution difference between different noise scenarios has been mentioned in traditional VADs. For example, in [16], Chang *et al.* used different statistical models for modeling the speech and noise distributions in different noise scenarios. Another related topic is the online learning methods [17]. they update the model parameters according to the historical domain information of the speech signals. Traditional statistical-model-based VADs [16] can also be regarded as unsupervised online learning methods. But to our knowledge, how to combine multiple features effectively is still an open problem in the online learning methods. On the other side, although our domain-adaptation-based VADs work in batch mode, it can combine multiple features effectively and also does not require heavy manual labeling.

The proposed methods are strongly related to the machine-learning-based speech separation and enhancement techniques [18–20]. The authors have used a large amount of noise scenarios to train the models and tested the models in the noise scenarios that have never appeared in the training set. Finally, they achieved significant performance improvement on the unseen noise scenarios. However, to our knowledge, the unsupervised domain adaptation techniques have not been considered in these works.

Table 1. Features and their attributes. The index of each feature is the window length of the feature [22].

ID	Feature	Dimension	ID	Feature	Dimension
1	Pitch	1	7	MFCC$_{16}$	20
2	DFT	16	8	LPC	12
3	DFT$_8$	16	9	RASTA-PLP	17
4	DFT$_{16}$	16	10	AMS	135
5	MFCC	20		**Total**	273
6	MFCC$_8$	20			

4. EXPERIMENTS

4.1. Experimental Settings

Seven noisy test sets of AURORA2 [21] is used for performance analysis. The signal-to-noise ratio (SNR) level of the audio signals is set to 5 dB. Each test corpus of AURORA2 contains 1001 utterances, which are split randomly into three groups for training, developing and test respectively. Each training set and development set consist of 300 utterances respectively, totally about 500 seconds long. Each test set consists of 401 utterances.

The sampling rate is 8kHz. We set the frame length to 25ms long with a frame-shift of 10ms. We extract 10 acoustic features from each observation. The detailed information of the features is listed in Table 1. All features are normalized into the range of $[0, 1]$ in dimension.

To simulate the real-world domain adaptation task, we take the training sets of the Street and Subway noise scenarios as two source corpora. For each source corpus, we form 6 domain adaptation tasks by randomly extracting a 30-second audio segment from the training set of each noise type of AURORA2, except that of the source corpus. For each domain adaptation task, the development set of the *source noise scenario* is used for model selection. We run each domain adaptation task 5 times and report the average accuracies. Note that the size of the target corpus is much smaller than the size of the source corpus.

The parameters are set as follows. Each deep model has three hidden layers. Only the best performance over all layers is reported. The numbers of the hidden units are set to $[54, 7, 7]$ respectively. The learning rate of the unsupervised pretraining is set to 0.004. The maximum epoch of the unsupervised pretraining is set to 200. The learning rate of the supervised fune-tuning is set to 0.005. The maximum epoch of the supervised fune-tuning is set to 130. The batch mode training is adopted. Each batch contains 512 observations.

To evaluate the effectiveness of the proposed domain adaptation schemes, we give the empirical lower bound and upper bound of the schemes. The **Lower Bound** is obtained by training DDNN with only the training set of the source noise scenario and testing it on various target noise environments. If the performance of the proposed domain adaptation schemes is worse than Lower Bound, it means that the schemes fail. The **Upper Bound** is obtained by training DDNN with the training set of the target noise scenario and testing the model on the test set of the same target environment. We also compare with the G.729B VAD [23], ETSI advanced frontend via Wiener filter [24], ETSI advanced frontend via frame dropping [24], Sohn VAD [25], Ramirez05 VAD [22], Ramirez07 VAD [26], Yu VAD [7], Shin VAD [8], and Ying VAD [17]. The experimental settings are exactly as those in [10].

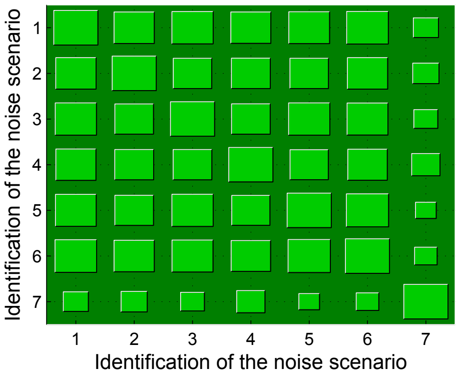

Fig. 1. Hinton diagram of the feature distributions in different noise scenarios. The identifications "1" to "7" represent babble, car, restaurant, street, airport, train, and subway respectively. Each grid of the Hinton diagram measures the distribution similarity of the features of the relevant two scenarios. The bigger the grid is, the more similar the two distributions are. The similarity is calculated by $\exp\left(-\|\mathbf{c}^{(s)} - \mathbf{c}^{(t)}\|^2/2\right)$ where $\mathbf{c}$ represents the feature centroid.

4.2. Experimental Results:

First, we show the Hinton diagram of the feature distributions of different noise scenarios in Fig. 1. From the figure, we can see that most feature distributions are relatively similar with each other except the subway noise scenario, which means domain adaptation might be useful.

Table 2 lists the accuracy comparison when the street noise is used as the source noise scenario. From the figure, we can see clearly that Scheme 2 is the most powerful one, followed by Scheme 1. But Scheme 3 is not effective. Moreover, Scheme $3^{(t)}$ is not only worse than Schemes 1 and 2, but also sometimes worse than the Lower Bound. This phenomenon manifested empirically that the domain adaptation schemes affect the performance significantly.

Table 3 lists the accuracy comparison when the subway noise is used as the source noise scenario. Because the performance is no better than the best referenced methods, we only

Table 2. Accuracy (in percentage) comparison when the street noise corpus (identification = 4) is used as the source corpus. Due to the length limit, we only report the best performance of the referenced VADs and its corresponding VAD algorithm. The referenced methods that are marked with "*" means that they are machine-learning-based VADs that are trained and tested in the matching environments. The line "Average improvement over Lower Bound" is calculated by $\frac{\text{Scheme }\#-\text{Lower Bound}}{\text{Upper Bound}-\text{Lower Bound}}\%$.

ID	Noise Type	Referenced	Scheme 1	Scheme 2	Scheme $3^{(s)}$	Scheme $3^{(t)}$	Lower Bound	Upper Bound
1	Babble	75.51 (Ramirez05)	**77.15**	76.59	75.73	73.17	74.95	79.14
2	Car	79.25 (G.729B)	82.91	**83.51**	82.92	82.20	82.05	87.09
3	Restaurant	69.59 (Ramirez05)	75.34	**75.74**	75.19	**75.76**	74.44	83.78
5	Airport	72.45 (Shin)*	**77.92**	**77.88**	77.51	77.32	77.35	82.30
6	Train	75.26 (G.729B)	81.69	**82.37**	81.42	80.88	80.51	84.25
7	Subway	73.16 (Ramirez05)	74.49	**76.42**	70.70	68.26	68.44	87.09
Average improvement over Lower Bound			25.79	30.88	13.93	-2.84		

Table 3. Accuracy (in percentage) comparison when the subway noise corpus (identification = 7) is used as the source corpus.

ID	Noise Type	Referenced	Scheme 1	Scheme 2	Scheme $3^{(s)}$	Scheme $3^{(t)}$	Lower Bound	Upper Bound
1	Babble	**75.51** (Ramirez05)	54.60	68.11	54.59	54.58	54.58	79.14
2	Car	**79.25** (G.729B)	58.05	70.05	63.09	64.11	61.33	87.09
Average improvement over Lower Bound			-6.33	44.47	3.44	5.40		

Table 4. Pretraining time (in seconds) comparison.

Scheme 1	Scheme 2	Scheme 3		
		Source	Target	Hybrid of top layer
774.87	12838.95	12592.76	570.35	985.92

show the results on the first two target noise scenarios without further running the remaining 4 tasks. Because the subway noise and the target noise scenario are significantly different, and also because the source corpus is much larger than the target corpus, from the table, we can see that the accuracies of all schemes drop significantly from Upper Bound. However, we can still observe that the accuracies from Scheme 2 are still significantly better than the Lower Bound.

Table 4 lists the pretraining time of the schemes. From the table, we observe that Scheme 2 is the slowest scheme. Although the source DDNN pretraining of Scheme 3 is slow, it needs to run only once, hence, Scheme 3 is still efficient.

Summarizing the aforementioned, Scheme 2 is the most effective one in dealing with the mismatching problem; although Scheme 3 is a failed scheme, it provides some useful information for our future work.

5. CONCLUSIONS

In this paper, we have tried to solve the mismatching problem between the training corpus and the test corpus via three DDNN-based domain adaptation schemes. To our knowledge, this is the first work that uses the powerful deep neural network to deal with the mismatching problem of VAD. Experimental results have shown that Schemes 2 is promising in dealing with the mismatching problem of VAD. The results also have shown that the layer-wise pretraining strategy has a significant impact on the deep-learning-based VADs. Although Scheme 3 is failed, it does provide an attempt to balance the training time and accuracy, and provide a contrary example for showing the effectiveness of the layer-wise pretraining. Experimental results also have shown that when the source and target corpora are very dissimilar, the performance is weaker than the referenced methods.

In the future, we will improve the performance of the unsupervised domain adaptation by training large-scale deep models with a large number of noise types using a massive computing power. We will also focus on the VAD problem in very low SNR environments, such as SNR = -5dB.

6. REFERENCES

[1] D. Yu and L. Deng, "Deep-structured hidden conditional random fields for phonetic recognition," in *Proc. INTERSPEECH*, 2010, pp. 2986–2989.

[2] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 30–42, 2012.

[3] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath

et al., "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 11, pp. 2–17, 2012.

[4] W. Hartmann, A. Narayanan, E. Fosler Lussier, and D. Wang, "A direct masking approach to robust ASR," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 1993–2005, 2013.

[5] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, "Applying support vector machines to voice activity detection," in *Proc. Int. Conf. Signal Process.*, vol. 2, 2002, pp. 1124–1127.

[6] Q. H. Jo, J. H. Chang, J. W. Shin, and N. S. Kim, "Statistical model-based voice activity detection using support vector machine," *IET Signal Process.*, vol. 3, no. 3, pp. 205–210, 2009.

[7] T. Yu and J. H. L. Hansen, "Discriminative training for multiple observation likelihood ratio based voice activity detection," *IEEE Signal Process. Lett.*, vol. 17, no. 11, pp. 897–900, 2010.

[8] J. W. Shin, J. H. Chang, and N. S. Kim, "Voice activity detection based on statistical models and machine learning approaches," *Computer Speech & Language*, vol. 24, no. 3, pp. 515–530, 2010.

[9] Y. Suh and H. Kim, "Multiple acoustic model-based discriminative likelihood ratio weighting for voice activity detection," *IEEE Signal Process. Lett.*, vol. 19, no. 8, pp. 507–510, 2012.

[10] X. L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 697–710, 2013.

[11] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, 2013, pp. 7378–7382.

[12] P. Teng and Y. Jia, "Voice activity detection using convolutive non-negative sparse coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, 2013, pp. 7373–7377.

[13] P. De Leon and S. Sanchez, "Voice activity detection using a sliding-window, maximum margin clustering approach," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, 2013, pp. 6674–6678.

[14] X.-L. Zhang and J. Wu, "Denoising deep neural networks based voice activity detection," in *Proc. 38th IEEE Int. Conf. Acoustic, Speech, Signal Process.*, 2013, pp. 1–5.

[15] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[16] J. H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, 2006.

[17] D. Ying, Y. Yan, J. Dang, and F. Soong, "Voice activity detection based on an unsupervised learning framework," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2624–2644, 2011.

[18] K. Han and D. L. Wang, "Towards generalizing classification based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 166–175, 2013.

[19] Y. X. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. PP, no. 99, pp. 1–23, 2013.

[20] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 134, pp. 3029–3038, 2013.

[21] D. Pearce, H. Hirsch *et al.*, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ICSLP'00*, vol. 4, 2000, pp. 29–32.

[22] J. Ramírez, J. C. Segura, C. Benítez, L. García, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 689–692, 2005.

[23] A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin, and J. P. Petit, "ITU-T Recommendation G. 729 Annex B: a silence compression scheme for use with G. 729 optimized for V. 70 digital simultaneous voice and data applications," *IEEE Commun. Mag.*, vol. 35, no. 9, pp. 64–73, 1997.

[24] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI ES*, vol. 202, no. 050.

[25] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[26] J. Ramírez, J. Segura, J. Górriz, and L. García, "Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2177–2189, 2007.