Multi-Task Adversarial Network Bottleneck Features for Noise-Robust Speaker Verification

Hong Yu1,2, Tianrui Hu2, Zhanyu Ma2*, Zheng-Hua Tan3, Jun Guo2

1Ludong University, Shandong, China
2Beijing University of Posts and Telecommunications, Beijing 100876, China
3Aalborg University, Denmark
hongyu@bupt.edu.cn, hutr96@126.com, mazhanyu@bupt.edu.cn, zt@es.aau.dk, guojun@bupt.edu.cn

Abstract: Modern automatic speaker verification (ASV) systems need to be robust under various noisy conditions. Motivated by the success of generative adversarial networks (GANs), this paper proposes a multi-task adversarial network (MAN) for extracting noise-invariant bottleneck (BN) features. The MAN consists of three component networks, a feature encoding network (FEN), a speaker discriminative network (SDN) and a noise-domain adaptation network (NAN). The FEN aims to generate noise-robustness BN features, the SDN makes the features from the FEN more speaker-discriminative and the NAN guides the FEN to learn more noise-invariant feature representations. The MAN is trained using an adversarial method. When training FEN and SDN, speaker identities and the label of being clean speech are used as target labels, which can make BN features, extracted from noisy or clean speech, similar. When training NAN, on the contrary, noise types are used as training targets. We evaluate the newly proposed MAN-BN feature extraction method on a Gaussian mixture model-universal background model (GMM-UBM) based ASV system. The experimental results on the RSR2015 database show that the proposed MAN-BN feature can dramatically improve the accuracy of the ASV system under different noise-type and signal-to-noise-ratio conditions.

Keywords: Speaker Verification, Multi-task Adversarial Training, Bottleneck Features

1 Introduction

The accuracy of classical automatic speaker verification (ASV) systems will be heavily decreased by adding noises. During test, background noise, particular unknown types of noise, will greatly degrade the performances of Gaussian mixture model universal background model (GMM-UBM) [1] and i-Vector methods [2] based ASV systems [3-4].

Many researchers have worked on developing noise robustness ASV systems during the last decade. In back-end solutions, multi-condition training methods pool clean and noisy speech together to train ASV models that make the trained systems fit the noisy conditions better [5-6].

In front-end solutions, some signal-to-noise ratio (SNR) increasing methods are selected to recover the clean speech from the noisy one, e.g., Wiener filter [7], short-time spectral amplitude minimum mean square error (STSA-MMSE) [8], and non-negative matrix factorization [9]. In the past few years, many deep learning methods are also used to generate clean speeches from noisy ones. In order to improve noise robustness of speech processing systems, in [10-12], regression methods are used to transform noisy speech features into clean features. In [13-14], noisy features and corresponding cleaning features are used to learn the ideal time-frequency binary mask (ITFBM) or ideal time-frequency ratio mask (ITFRM) and the estimated masks are then used to recover clean speech.

Recently, a generative adversarial network (GAN) [15] was proposed and used in many signal generation tasks, such as image generation [16] and image to image translation [17-19]. Besides image generation tasks, adversarial training ideas are also recently used in the speech processing domain for phone/senone classifiers [20-21], music generation [22] and speech enhancement [23].

In our previous work [24], we use a single task adversarial network to generate noise-robust features and the method only focuses on noise adaptation without considering speaker discriminative information which is very important to the ASV task. In this paper, we build a multi-task adversarial network (MAN) to generate noise-invariant speech representation including both the speaker-discriminative part and the noise-robustness part. The bottleneck (BN) features produced by MAN both include speaker distinguish information and noise invariant attribute.

In Section 2, we introduce the structure of the proposed MAN and describe the details of training method. Five baseline front-ends used for comparison purpose are introduced in Section 3. In Section 4, we introduced the RSR2015 database and noise data which are used for neural network training and the ASV performance evaluation. In Section 5, we compare the performance of proposed front ends with six other speech enhancement based and neural network bottleneck feature based front-ends on the ASV system, and the conclusion is drawn in Section 6.

2 MAN-BN feature extractor
Recently, GANs [15] has got a lot of attention and has been used in many image and speech generation tasks. A GAN is composed of two networks: a generator network (\(G\)) and a discriminator network (\(D\)). \(G\) can generate ‘fake’ samples from random inputs, \(z\), \(D\) is a classifier with a sigmoid output, which can distinguish the ‘fake’ sample (\(G(z)\)), from the ‘real’ sample, \(x\). An adversarial training method is used to update parameters \(\theta_G\) in \(G\) and \(\theta_D\) in \(D\) in turn following a min-max game with the objective as shown in equation (1) as follows:

\[
V(D,G) = E_{x \sim p_{data}(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log (1 - D(G(z))]
\]

(1)

Firstly, we fix \(\theta_D\) to maximize \(V(D,G)\) by updating \(\theta_D\) in order to make the discriminator \(D\) better distinguish between ‘real’ and ‘fake’ samples. Then, we fix \(\theta_D\) to minimize \(V(D,G)\) by updating \(\theta_G\) to make the ‘fake’ data generated by \(G\) more similar to the ‘real’ one.

If the discriminator is chosen to have two softmax outputs, the label of ‘real’ samples are set as \(I_{\text{real}} = [1, 0]\) and the label of ‘fake’ samples are set as \(I_{\text{fake}} = [0, 1]\). We use stochastic gradient descent to update \(\theta_G\) and \(\theta_D\). Then, minimizing and maximizing equation (1) can be changed to minimization of two cross-entropy (CE) cost functions:

\[
\min_{\theta_G} CE_G = \min_{\theta_G} \frac{1}{M} \sum_{i=1}^{M} [I_{\text{real}} \log D(x^{(i)}) + I_{\text{fake}} \log D(G(z^{(i)}))]
\]

(2)

\[
\min_{\theta_D} CE_D = \min_{\theta_D} \frac{1}{M} \sum_{i=1}^{M} [-I_{\text{fake}} \log D(G(z^{(i)}))]
\]

(3)

Where \(M\) stands for the number of training samples. As suggested in paper [15], we usually use equation (4) to replace equation (3).

\[
\min_{\theta_G} CE_G = \min_{\theta_G} \frac{1}{M} \sum_{i=1}^{M} [I_{\text{real}} \log D(G(z^{(i)}))]
\]

(4)

It can be seen clearly that when updating \(D\), the goal is to classify ‘real’/‘fake’ data and when training \(G\), the goal is to make the generated ‘fake’ sample similar to the ‘real’ one (by using the ‘real’ target label).

Inspired by the structure of GAN, we build a MAN by changing the random input \(z\) as real acoustic features and changing the single task generator to a multi-task generator.

The proposed MAN-BN feature extractor consists of three joint networks, a feature encoder network (FEN), a speaker discriminative network(SDN) and a noise adaptation network(NAN), as shown in Figure 1. In FEN, there are three hidden layers, FE1, FE2, and FE3, where FE1 and FE2 have 1024 nodes and the feature extraction layer FE3 has 128 nodes. The activation functions of FE1 and FE2 are softplus \((\log(\exp(x) + 1))\). In FE3, tanh is selected as the activation function, the output of FE3 is used as the MAN-BN feature.

The SDN includes two sigmoid hidden layers with 1024 nodes each and a softmax output layer. The dimension of the output layer is \(N_{pk}\), representing \(N_{pk}\) training speakers. Same as SDN, NAN also includes two 1024 sigmoid hidden layers and a softmax output layer. The dimension of the output layer is \(N_{noise} + 1\), the first dimension stands for clean speech and the other \(N_{noise}\) dimensions represent different noise types. During the MAN training, our target is to obtain MAN-BN features that are good for speaker recognition but harmful for noise type classification. Meanwhile, the NAN should also have a good noise type classification accuracy. So we design two cross entropy cost functions for the multi-task adversarial training.

\[
\min_{\theta_{FE}, \theta_{SD}, \theta_{NA}} CE_1 = \min_{\theta_{FE}, \theta_{SD}, \theta_{NA}} \frac{1}{M} \sum_{i=1}^{M} [I_{\text{real}} \log SP(\text{FE}(x^{(i)})) + \beta \log \text{NA}(\text{FE}(x^{(i)}))]
\]

(5)

\[
\min_{\theta_{NA}} CE_2 = \min_{\theta_{NA}} \frac{1}{M} \sum_{i=1}^{M} [I_{\text{noise}} \log \text{NA}(\text{FE}(x^{(i)}))]
\]

(6)

Where \(\theta_{FE}, \theta_{SD}, \theta_{NA}\) and \(\text{FE}, \text{SD}, \text{NA}\) stand for parameters and forward transfer function of FEN, SDN and NAN, respectively. \(l_{\text{real}}\) stands for the one-hot speaker-id target labels and \(l_{\text{noise}}\) means the label of clean speech. The first term of equation (5) can make the MAN-BN feature generated from FEN helpful for speaker classification and the second term of CE1, can make MAN-BN noise invariant by using the same clean speech label, \(l_{\text{real}}\) as training targets, no matter the input \(x^{(i)}\) is clean or noisy feature. \(\alpha\) and \(\beta\) are two weight parameters that can balance the weight between speaker discrimination and noisy adaptation. In equation (6), we use real noise types, \(l_{\text{noise}}\), as training target, which can help NAN to have a good noise classification ability.

In the training phase, firstly, we fix \(\theta_{SD}\) and update \(\theta_{FE}\) and \(\theta_{NA}\) by minimizing CE1 using a gradient descent method. Then we fix \(\theta_{SD}\) and \(\theta_{NA}\) to minimize CE2 by updating \(\theta_{FE}\). Execute these two steps in turn until two cost functions become convergence. Then we can use the trained FEN to generate noise invariant features to improve the performance of ASV systems under noisy conditions.

3 Baseline systems

In this section, we introduce five baseline front-ends, STSA-MMSE, deep neural network speech enhancement (DNN-SE), speaker discriminative network bottleneck (SDN-BN) feature, single task adversarial network bottleneck features (SAN-BN) described in paper [24] and feature generated by the domain adaptation network (DAN) introduced in paper [20] [25]. We also describe
the ASV baseline system which will be used to evaluate the performances of different front-ends.

3.1 STSA-MMSE

STSA-MMSE is a speech enhancement method which based on the assumption that discrete Fourier transform (DFT) coefficients of noise free speech follow a generalized gamma distribution [8]. Because it does not need the priori knowledge of noise type or level, it is widely used in many speech preprocessing area. In our experiment, the priori SNR is estimated by the Decision-Directed approach [26] and the noise power spectral density (PSD) is estimated by the noise PSD tracker reported in [27] which is estimated by the first 1000 samples.

3.2 DNN based speech enhancement

Ideal ratio mask (IRM) estimation based DNN-SE method introduced in [13] is used as another baseline. As suggested in [13], IRM used for training targets are generated by a gammatone filter banks in frequency-domain and in time-domain the output of each filter bank channel is divided into 20 ms frames with 10 ms overlap.

We uses a DNN with three 1024 sigmoid hidden layers to estimate the IRM of input noisy speeches. The DNN is trained by acoustic features with 1845 dimensions and mean square error is selected as the cost function. The estimated IRM of testing noisy speeches is used to reconstruct the time-frequency representation of enhanced speech and recover the clean speech from noisy one.

3.3 SDN-BN

Many researches show that BN features generated from speaker discriminative networks can improve the performance of ASV systems [28-29]. So in this paper, we also evaluate the performance of SDN-BN features on noisy conditions. In order to make a fair comparison with the MAN-BN feature introduced in this paper, we delete the NAN part of MAN, and the rest part is trained as a speaker discriminative network, using speaker-ids as target labels. The output of FEN is used as SDN-BN feature to train ASV systems.

3.4 SAN-BN

In paper [24], a single task adversarial network (SAN) is designed to generate noise-robust acoustic features. Unlike the MAN described in this paper, both speaker recognition and noise adaptation tasks are realized in one network. The SAN in [24] only focuses on the noise adaptation task, which means the SDN part of Fig 1 is deleted, and the first term of equation (5) is also omitted. Only \( l_s \) and \( l_n \) are used as training targets to update FEN and NAN. Outputs of FEN are also used to train ASV systems.

3.5 DAN-BN

In paper [25] and [20], a DAN is introduced and used on image classification and speech recognition tasks. The network has the similar structure to the one shown in Fig 1. While in the training phase, the cost function used to update the network is designed as equation (7).

In the third term of equation (7), the author uses a negative target label, \(-l_s\) to realize the noise adaptation function. When using these negative target labels to update FEN, a gradient reversal layer was used to multiplies the gradient from the negative target label, during the back propagation based training. Here, we also use bottleneck outputs of FEN, as features for ASV system training.

\[
\min \frac{\sum}{m} CE = \min \frac{1}{M} \sum_{m} \left( \begin{array}{c}
\alpha_s \log \frac{\text{SP}(\text{FE}(x^{(m)}))}{\text{NA}(\text{FE}(x^{(m)}))} \\
\alpha_n \log NA(\text{FE}(x^{(m)})) - \beta \end{array} \right)
\]

3.6 ASV systems

Classical GMM-UBM based speaker verification systems are used to evaluate the noise robustness of six different features. The GMM-UBM method is chosen as it performs well for short utterances [28] [30], which is the case in this paper. A large quantity of speaker in-depended speeches are used to train a universal background GMM model (UBM), firstly. Then, we use maximum-a-posteriori (MAP) adaptation method to build GMMs of enrollment speakers. In the testing phase, the log-likelihood ratio between the claimed speakers’ GMM and the UBM are used as speaker verification scores.

4 Speech corpora and noise data

Same as the setting in paper [24], we use all male speaker utterances in TIMIT corpus [31] to train the UBM. Clean speeches used for training MAN, SAN, DAN, SDN-BN and speaker models and for testing ASV are all from RSR2015 corpus [32] as detailed in Table I.

| Table 1 Trails used for different front-ends training, speaker models training and ASV testing |
|---|---|---|---|
| SDN | Text ID | Sess.ID | Spk.ID |
| SAN | 2-30 | 1,4,7 | 51-100 |
| MAN | | | |
| DAN | | | |
| DNN-SE | | | |

| Spk Model | Text ID | ASV test | |
|---|---|---|---|
| ASV test | 1 | 2,35,6,8,9 | 2-50 |

We build a text-dependent ASV system to evaluate the performance of six different noise robustness features. As shown in Table 1, we use text ID 1 and three sessions (1,4,7) of 49 male speakers (m002 to m050) to train speaker models. In testing phase, sessions 2, 3, 5, 6, 8, and 9 are used.

In neural networks training steps, SDN, SAN, MAN, DAN, and DNN-SE models are all trained using the same utterances of text IDs 2-30 and sessions 1,4 and 7 from 50 male speakers (m051-m100). Speeches used for neural networks, speaker model training and ASV testing are recorded by different devices which can make an unmatched training/testing condition.

In this paper, we consider five different types of noise: Babble, Cantile, Market, Airplane and white Gaussian.
noise (White). White was generated in MATLAB, Babble was produced fusion 6 randomly selected speeches in Librispeech database [33], Cantine, Market and Airplane noises were collected by the author and Fondazione Ugo Bordoni (FUB) [34]. All noise data are split into three non-overlapping parts, which are used in different front-end training, speaker model training and speaker verification testing, respectively.

5 Experimental results and discussion

We use clean and all kinds of noisy speeches to train five neural networks based front-ends. Noisy speech is generated by adding desired SNR levels noises on clean speeches, using the ITU toolkit [35] SNRs of noisy speech are selected as 10dB and 20dB.

DNN-SE is trained by 1845 dimension acoustic features [12], the other four front-ends are all trained by mel-frequency cepstral coefficients (MFCCs) with 57 dimensions. Inputs of SDN, SAN, MAN and DAN are batch normalized 11 sequential feature frames. The mixture of clean and noisy speeches are used as training data. Stochastic gradient descent (SGD) is used as the optimized method and the mini-batch size is set as 30.

For MAN and DAN training we set $\alpha = 0.7$ and $\beta = 0.3$. Because the noise discriminative part in MAN ($CE_2$) converges faster, in order to balance the updating of $CE_1$ and $CE_2$, in each mini-batch training, we update $CE_2$ with a 50% probability only.

In evaluation phase, we used MFCC based no enhancement front-end as baseline. In the baseline front-end and wave enhancement based STSA-MMSE and DNN-SE front-ends, we used MFCCs for ASV training and testing. For the SDN, SAN, MAN, and DAN front-ends, bottleneck features with 128 dimensions are used for GMM-UBM based speaker models training and testing. The mixture number of GMMs is chosen as 512.

The performance of different front-ends under different noisy level (from 0dB-20dB) in ASV systems is evaluated by equal error rates (EER). For baseline front-end, we only use clean speech to build the ASV system and for other front-ends, in order to build a matching training and testing condition, enhanced clean speech or bottleneck features generated by clean speech are used.

Table II Average EER(%) in different front-ends

|       | White | Babble | Cantine | Market | Airplane |
|-------|-------|--------|---------|--------|----------|
| No Enh. | 28.98 | 12.10  | 13.66   | 15.39  | 13.09    |
| MMSE  | 15.10 | 13.66  | 9.72    | 12.70  | 9.51     |
| DNN-SE | 15.57 | 8.73   | 8.69    | 9.58   | 8.08     |
| SD-BN | 12.54 | 6.43   | 5.17    | 5.47   | 4.47     |
| SAN-BN | 12.07 | 7.17   | 5.00    | 6.18   | 5.42     |
| DAN-BN | 12.47 | 6.65   | 4.95    | 5.73   | 4.48     |
| MAN-BN | 11.86 | 6.16   | 4.91    | 5.53   | 4.31     |

Average EERs under different noise levels are shown in Table II, it can be observed clearly that comparing with the speech enhancement based front-ends, e.g., MMSE and DNN-SE, bottleneck features based front-ends are more suitable for the ASV task. The average EER can be decreased by about 20%-40%.

Generally speaking, single task based front-ends, SD-BN and SAN-BN, which are focusing on speaker discrimination and noise adaption, respectively, perform worse than the multitask based front-ends. Comparing with the two multi-task adversarial network based front-ends, MAN-BN introduced in this paper works better than the DAN-BN. A reasonable explanation can be found in the different cost functions.

Refer to GAN cost functions, equation (3) and (4) in Section 2, equation (7) is corresponding to equation (3), using negative 'fake' labels to train $G$, while equation (5) is corresponding to equation (4), which is suggested, using 'real' labels to train the generator. During the MAN training, we use equation (5) to realize the noise adaption by giving a unified clean speech target label, $lc$, to every input. While in DAN training, as shown in equation (8), negative target labels, $-ln$, are used in order to damage the function of the noise type classifier.

The discriminative information of noise types is still included in negative target labels, which will cause the extracted bottleneck features do not meet the noise invariable condition. That is why in some situation, the DAN-BN feature performs worse than MAN-BN based front-ends.

6 Conclusions

A new multi-task adversarial networks (MAN) based noise robust feature extractor is introduced in this paper. The proposed MAN consists of three joint networks, a feature encoding network (FEN), a speaker discriminative network (SDN) and a noisy domain adaptation network (NAN). FEN, SDN and NAN are trained in turn and the outputs of the FEN are used as robust features for automatic speaker verification (ASV). When updating parameters of FEN and SDN, a multi-task cross entropy cost function is selected which using speaker-id and clean speech label as target labels, in order to make the output of FEN speaker discriminative and noise invariant. When training NAN, noise types are used as target labels to make NAN have a good noise classification ability. By using multi-task adversarial training, we can obtain MAN-BN features which are both noise robust and suitable for the ASV task. Comparing different performances of different front-ends, the newly proposed MANBN front-end outperforms the speech enhancement based frontends and some other deep neural network bottleneck features under noise conditions.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) No. 61773071, in part by the Beijing Nova Program No. Z17110000117049, in part by the Beijing Natural Science Foundation (BNSF) No. 4162044, in part by the Beijing Nova Program Interdisciplinary Cooperation under Project (Z181100006218137).

References
Proceedings of IC-NIDC 2018

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” Digital signal processing, vol. 10, no. 1-3, pp. 19–41, 2000.

[2] N. Dehak, P. J. Kenny, R. Dehak, and et al., “Front-end factor analysis for speaker verification,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788–798, 2011.

[3] F. Yan, A. Men, B. Yang, and Z. Jiang, “An improved ranking-based feature enhancement approach for robust speaker recognition,” IEEE Access, vol. 4, pp. 5258–5267, 2016.

[4] O. Plchot, S. Matsoukas, P. Matejka, and et al., “Developing a speaker identification system for the darpa rats project,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 6768–6772.

[5] Y. Lei, L. Burget, and N. Scheffler, “A noise robust i-vector extractor using vector taylor series for speaker recognition,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 6788–6791.

[6] Y. Lei, L. Burget, L. Ferrer, and et al., “Towards noise-robust speaker recognition using probabilistic linear discriminant analysis,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 4253–4256.

[7] P. Scalart et al., “Speech enhancement based on a priori signal to noise estimation,” in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 2, 1996, pp. 629–632.

[8] J. S. Erkenlis, and et al., “Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 6, pp. 1741–1752, 2007.

[9] N. B. Thomsen, D. A. L. Thomsen, Z.-H. Tan, and et al., “Speaker-dependent dictionary-based speech enhancement for text-dependent speaker verification,” Interspeech 2016, 2016.

[10] O. Plchot, L. Burget, H. Aronowitz, and and et al., “Audio enhancing with dnn autoencoder for speaker recognition,” in ICASSP 2016, 2016, pp. 5090–5094.

[11] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7–19, 2015.

[12] M. Kolbek, Z.-H. Tan, and J. Jensen, “Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification,” in Spoken Language Technology Workshop, 2016, pp. 305–311.

[13] M. Kolbek, Z.-H. Tan, and J. Jensen, “Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 1, pp. 153–167, 2017.

[14] 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.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in neural information processing systems, 2014, pp. 2672–2680.

[16] G.-J. Qi, “Loss-sensitive generative adversarial networks on lipschitz densities,” arXiv preprint arXiv:1701.06264, 2017.

[17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” arXiv preprint arXiv:1611.07004, 2016.

[18] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” arXiv preprint arXiv:1511.06434, 2015.

[19] J. Zhao, M. Mathieu, and Y. LeCun, “Energy-based generative adversarial network,” arXiv preprint arXiv:1609.03126, 2016.

[20] D. Serdyuk, K. Audhkhasi, P. Brakel, and et al., “Invariant representations for noisy speech recognition,” arXiv preprint arXiv:1612.01928, 2016.

[21] Y. Shinohara, “Adversarial multi-task learning of deep neural networks for robust speech recognition,” Interspeech 2016, pp. 2369–2372, 2016.

[22] O. Mogren, “C-rnn-gan: Continuous recurrent neural networks with adversarial training,” arXiv preprint arXiv:1611.09904, 2016.

[23] D. Michelsanti and Z.-H. Tan, “Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification,” Proc. Interspeech 2017, pp. 2008–2012, 2017.

[24] H. Yu, Z.-H. Tan, Z. Ma, and J. Guo, “Adversarial network bottleneck features for noise robust speaker verification,” Interspeech 2017, 2017.

[25] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” arXiv preprint arXiv:1409.7495, 2014.

[26] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 6, pp. 1109–1121, 1984.

[27] R. C. Hendriks, R. Heusdens, and J. Jensen, “Mmse based noise psd tracking with low complexity,” in IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2010, pp. 4266–4269.

[28] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, “Deep feature for text-dependent speaker verification,” Speech Communication, vol. 73, pp. 1–13, 2015.

[29] Z. Zhang, L. Wang, A. Kai, T. Yamada, and et al., “Deep neural network-based bottleneck feature and denoising autoencoder-based denoiser for distant-talking speaker identification,” EURASIP Journal on Audio, Speech, and Music processing, vol. 2015, no. 1, p. 12, 2015.

[30] H. Delgado, M. Todisco, M. Sahidullah, and et al., “Further Optimisations of Constant Q Cepstral Processing for Integrated Utterance and Text-dependent Speaker Verification,” in Processing of Spoken Language Technology Workshop (SLT), 2016.

[31] J. S. Garofolo, L. F. Lamel, W. M. Fisher, and et al., “Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1-1,” NASA STI/Recon technical report n, vol. 93, 1993.

[32] A. Larcher, K. A. Lee, B. Ma, and H. Li, “Text-dependent speaker verification: Classifiers, databases and RSR2015,” Speech Communication, vol. 60, pp. 56–77, 2014.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in ICASSP, 2015, pp. 5206–5210.

[34] M. Falcone, M. C. Benoit Fauve et al., “Corpora collection,” OCTAVE (Objective Control of TalkER VERIFICATION), Deliverable 17, 2016.

[35] G. Recommendation, “191: Software tools for speech and audio coding standardization,” Int. Telecommun. Union, Geneva, Switzerland, 2005.