Deep Neural Network Architectures for Double Talk Detection in AEC

Ziteng Wang[1], *Emmanuel Vincent*[2], *Xiaofei Wang*[1], *Yonghong Yan*[1]

[1]Institute of Acoustics, University of Chinese Academy of Sciences, China
[2]INRIA Nancy, France

wangziteng@hccl.ioa.ac.cn

Abstract

In Acoustic Echo Cancellation (AEC), the adaptive filter diverges from its optimum solution when double talk occurs. In this paper, the problem of Double Talk Detection (DTD) is addressed from a machine learning perspective. The Deep Neural Network (DNN) is introduced as a binary classifier, which is trained on spectral features of the far-end speech and the microphone signal. To further make use of the available data, a separate network is proposed to learn the echo path effect. The network, which maps the far-end speech to its echo image, actually collaborates with the classifier. Experiments are conducted to demonstrate its performance and the robustness to environment changes is analyzed.

Index Terms: double talk detection, deep neural network, echo path, acoustic echo cancellation

1. Introduction

In a full-duplex system as illustrated in Fig.1, Acoustic Echo Cancellation (AEC) is an indispensable module which is used to suppress the undesirable echo of the far-end speech. An adaptive filter is designed to identify the echo path and then the echo component is subtracted from the microphone signal. If both the far-end speaker and the near-end speaker talk simultaneously, the so-called double talk occurs. The near-end speech acts as uncorrelated noise to the adaptive algorithm and diverges the filter from its optimum solution. Therefore Double Talk Detection (DTD) is introduced to control the updating of the adaptive filter. When double talk is detected, the filter update is slowed down or completely halted [1].

The solution to the DTD problem is a robust discriminator. The discriminator should achieve a high detection accuracy while keeping a low level of false alarm probability, regardless of the ambient noise and the changes in the echo path. In literatures, the common routine is to design a feature or detection statistic from the available signals. By comparing its value to a heuristic threshold, a binary decision is made: double talk or no double talk. The classical Geigel algorithm [2] is based on an energy feature, which just involves the magnitudes of the recent past samples of the far-end speech. A generalized approach based on the Holder inequality is proposed in [3], which takes the effect of noise into consideration. In [4], a feature set is designed which includes the standard deviation and the log-energy of the signals. A large family of methods are based on the cross-correlation [5, 6, 7, 8, 9, 10] and the coherence [11, 12]. The correlation statistic is usually calculated between the residual echo signal and the microphone signal [5]. However, it is shown vulnerable to the near-end disturbance and a time delay is introduced to improve the performance [6]. Another option is the cross correlation between the far-end signal and the microphone signal [7]. The Normalized Cross-Correlation (NCC) methods are discussed in [8, 9]. There are also other methods based on the signal envelop [13] and the zero-crossing rate [14], which are intended for the cases of low computational capacity.

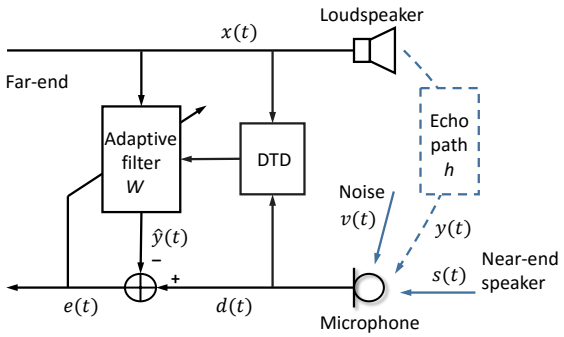

Figure 1: *AEC in a full-duplex system. The DTD module controls the updating of the adaptive filter.*

The methods above generally exploit no prior information of the acoustic environment. While in many scenarios such as in a conference room, the positions of the loudspeaker and the microphone are fixed over time, the acoustic properties of the environment can then help the detection of a double talk instance. In [15], the double talk decision is made with the estimated echo path. A known loudspeaker impulse response is assumed and utilized in [16]. Besides, the fact that data samples can be easily obtained in these scenarios motivates data-driven methods. A learning based approach is proposed in [17], which employs a one recurrent layer network. The GMM model is considered in [18] along with frequency domain features. Notably, the idea of deep learning already achieves success in many applications, such as in a similar task of Voice Activity Detection (VAD) [19, 20], though VAD is designed to distinguish between speech and noise periods. One AEC related work is the residual echo suppression with Deep Neural Network (DNN) in [21]. DNN shows better modeling ability and it is advantageous in dealing with the potential non-linearity.

In this paper, we apply the deep learning methodology to the DTD problem. The problem can be easily formulated as a binary classification task, for which DNN is introduced as a discriminative classifier. The available training samples can be further utilized by a separate network, which is proposed to learn the echo path effect. The far-end speech is mapped to its echo images. This network indeed collaborates with the classifier and benefits the performance. Especially, its functionality is different from the common practice of cancelling the echo effect as the network in [21]. And such a network is shown not proper for the task. Simulations are performed in the experiments and the behaviour of these networks is analyzed.

The outline of this paper is as follows. In section 2, the problem of DTD is defined. Section 3 introduces the proposed

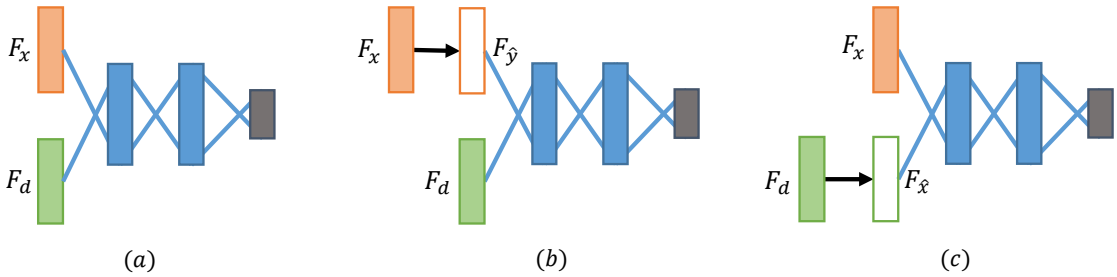

Figure 2: *The proposed DNN architectures: (a) the binary classifier (b) the classifier with an echo learning net (c) the classifier with an echo suppression net.* F_x, F_d, F_y *are respectively the feature vectors of the far-end signal, the microphone signal and the echo.*

DNN based classifiers. The evaluation results are given in section 4 and conclusions are drawn in section 5.

2. Double talk detection

As shown in Fig.1, the signal captured in the microphone is

$$d(t) = y(t) + s(t) + v(t) \tag{1}$$

where t is the time index, $s(t)$ is the near-end speech, $v(t)$ is the ambient noise, and $y(t)$ is the echo of the far-end signal $x(t)$. $y(t) = h_L * h(t) * x(t)$ with $*$ denoting convolution and $h_L, h(t)$ being the loudspeaker response and the echo path response. The adaptive filter $\boldsymbol{W}$ is designed to identify the system response, so the echo can be estimated and then subtracted from the microphone signal. The residual signal is

$$e(t) = d(t) - \boldsymbol{W}^T \boldsymbol{x} \tag{2}$$

where $\boldsymbol{x} = [x(t)\ x(t-1)\ ...\ x(t-L+1)]^T$ and ^{T} denotes transpose.

When double talk occurs, $\boldsymbol{W}$ will diverge from its optimal value and tend to cancel the near-end speech. To detect double talk, two hypothesises are made

$$H_0 : s(t) = 0, \quad \text{no double talk} \tag{3}$$
$$H_1 : s(t) \neq 0, \quad \text{double talk} \tag{4}$$

Then a general inference procedure is

1) Design a detection statistic ξ with the available signals, e.g. $x(t), d(t)$ and $e(t)$.

2) Compare ξ to a preset threshold and make a binary decision: either H_0 or H_1.

3) Once H_1 is declared, the status is held for a minimum period and the filter update is disabled.

As an example, the NCC method in [9] is based on the following statistic

$$\xi_{NCC} = 1 - \frac{r_{ed}}{\sigma_d^2} \tag{5}$$

where $r_{ed} = \mathbb{E}[e(t)d(t)]$ and σ_d^2 is the variance of $d(t)$. $\mathbb{E}$ denotes expectation. The theoretical decision threshold is 1. If $\xi_{NCC} < 1$, then double talk occurs.

3. The DNN architectures

In the following, the DTD problem is considered as a binary classification task, which can be solved by learning from examples. DNN is employed to perform supervised training. The proposed network architectures are illustrated in Fig.2.

3.1. Features and targets

The input features are the logarithm of the filter-bank energy extracted from the far-end signal and the microphone signal. The calculation requires Short-time Fourier Transform (STFT), which in practice can be accomplished easily together with frequency domain AEC. And the frame shift is adjusted accordingly to give a timely response. The difference between the two signal features is also used. Suppose there are K filter bands, then the input vector is of $3K$ in size, which is denoted as $\{\boldsymbol{F}_x, \boldsymbol{F}_d, \boldsymbol{F}_{x-d}\}$.

The output targets are defined by two vectors: $[1, 0]^T$ for H_0 and $[0, 1]^T$ for H_1. Two nodes are used rather than one so that discriminative training can be performed.

3.2. Binary classifier

DNN is mainly employed as a binary classifier as in Fig.2(a). It is a basic feed-forward network with all the layers linearly connected. The activation function is ReLU for the hidden layers and Softmax for the last layer. Cross-entropy is used as the loss function:

$$loss_{CE} = -\sum [p\log(\hat{p}) + (1-p)\log(1-\hat{p})] \tag{6}$$

3.3. Echo learning

The training data, specifically the date samples of class H_0, is further utilized in a separate echo learning network. A mapping from $\boldsymbol{F}_x$ to the echo image $\boldsymbol{F}_y$ is learned as in Fig.2(b) or an inverse mapping is learned as in Fig.2(c). A test sample is first processed by the mapping network and then sent to the classifier. We now investigate the features actually used for classifying the two classes. In the former case, the features are

$$H_0 : \begin{Bmatrix} \boldsymbol{F}_{\hat{y}} \\ \boldsymbol{F}_{y+v} \end{Bmatrix} \quad H_1 : \begin{Bmatrix} \boldsymbol{F}_{\hat{y}} \\ \boldsymbol{F}_{y+s+v} \end{Bmatrix}$$

where the difference feature vectors are omitted. Here the echo path effect and the nonlinearity introduced by the loudspeaker are partially handled by the echo learning network. An easier classification task is expected with a better learned mapping. In the latter case, the network learns to cancel the echo effect, which is a process commonly seen in echo suppression tasks. The features become

$$H_0 : \begin{Bmatrix} \boldsymbol{F}_x \\ \boldsymbol{F}_{\hat{x}+v'} \end{Bmatrix} \quad H_1 : \begin{Bmatrix} \boldsymbol{F}_x \\ \boldsymbol{F}_{\hat{x}+s'+v'} \end{Bmatrix}$$

Now a well trained echo suppression network should be robust to the noise and the near-end interferences, which may lead to a

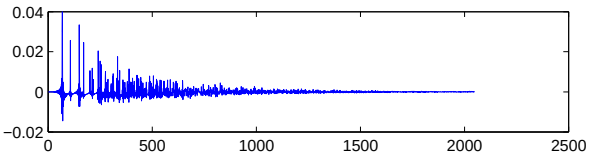

Figure 3: *The simulated echo path response.*

blurred decision surface between the two classes. The behavior of the two networks is shown in the experiments. During training, both mapping networks are tuned under the Mean Squared Error (MSE) criterion:

$$loss_{MSE} = \sum |\hat{\boldsymbol{F}} - \boldsymbol{F}|^2 \tag{7}$$

4. Experiment and analysis

4.1. Setup

The experimental data comes from the TIMIT corpus. The speech sentences are sampled at 16 kHz. The echo path response is simulated in a room with reverberation time 230 ms. It is truncated to 2048 taps and shown in Fig. 3. A moderate nonlinearity effect is considered in the loudspeaker, which is approximated with a memoryless sigmoidal function [22]

$$z = \frac{2}{1 + \exp(-a \cdot b)} - 1 \tag{8}$$

where

$$b = 1.5 \cdot x(t) - 0.3 \cdot x^2(t) \tag{9}$$

and the slop variable a is set 2 for $x > 0$ and 1 for $x < 0$.

The training and testing data is separately drawn from the train set and the test set. There are 2000 sentences for class H_0 and 800 sentences for H_1 since double talk instances are less frequent than no double talk instances. 160 sentences for each class are used as the development set. White Gaussian noise is added to the echo signal at Signal-to-Noise Ratio (SNR) of 30 dB. In the double talk case, the Near-end-to-Echo Ratio (NER) is set 0 dB. 100 sentences for each class are chosen for the final performance evaluation.

STFT is performed in 512 points with a 25 ms frame length and a 10 ms frame shift. 26 filter bands are used and only one frame is used as input. Then the binary classifier is 78-512-512-2 in size and the echo learning net is 26-512-512-26 in size. The mapping network is first trained and then fixed as a preprocesser for the classifier. In the training phase, the weight matrixes are initialized with i.i.d Gaussian samples and the bias vectors are filled with zeros. The Adam method is used for tuning the networks. Early stopping is applied when the loss in the development set no longer decreases. In the testing phase, the outputs of the classifier represent probabilities.

The metrics proposed in [7] are used for evaluation: the probability of detection P_d and the probability of false alarm P_f. In general, there exists a tradeoff between the two metrics and a higher P_d is achieved at the cost of a higher P_f.

4.2. Performance results

The performance of the three networks proposed in section 3 are given by the Receiver Operating Characteristic (ROC) in Fig.4. ROC is a curve of P_d as the function of P_f. The Area Under the Curve (AUC) values are shown in the legends. A higher AUC indicates better performance. The three networks in Fig.2 are separately denoted as Anet, Bnet and Cnet. The NCC method [9] is included for comparison. An adaptation algorithm is needed in the NCC statistic calculation, for which the classical Normalized Least Mean Squares (NLMS) [23] is used. The filter $\boldsymbol{W}$ is updated as

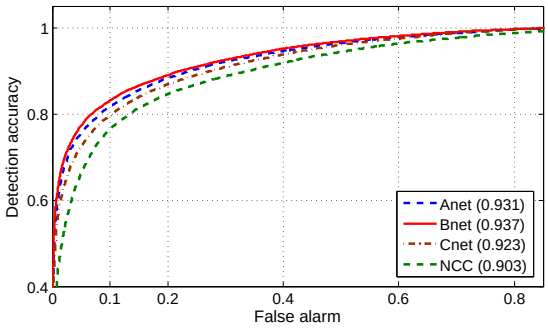

Figure 4: *The ROC curve of the classifiers. The AUC values are shown in the legend.*

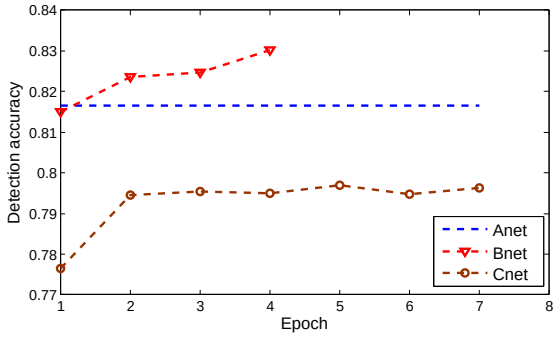

Figure 5: *The performance of the classifiers under $P_f = 0.1$ as the development error decreases.*

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t + \mu \frac{e(t)\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x} + \delta} \tag{10}$$

where $\mu = 0.4$ and $\delta = 0.001$ for numerical stability.

All the three networks outperform the NCC method and the Bnet performs best as expected. At a false alarm probability of 0.10, the basic classifier Anet achieves a 0.05 point higher accuracy than NCC. The Bnet is even 0.02 point better while the Cnet slightly deteriorates the performance of the classifier. The AUC scores of the networks are comparative, but still a difference is observed.

To support the arguments in subsection 3.3, the performance of the classifiers in each epoch is shown in Fig.5. The echo learning net is first trained. For each net with a decreased development error, it is saved and then a succeeding classifier is trained. For the Bnets, the detection accuracy goes up as the development error decreases. For the Cnets, though the development error decreases with more training epochs, the detection performance remains at a stable level.

4.3. Environment robustness

Note that the networks are trained in one single condition, they are tested for their generalization ability. The environment robustness test covers the near-end speech level, the ambient noise

level and the variations in the echo path. The NER is changed from 0 dB to {10, 5, -5} dB, and the SNR is changed from 30 dB to {35, 25, 20} dB. The echo path variations are simulated by slightly changing both the position of the loudspeaker and the Reverberation Time (RT). The results of detection accuracy under a maximum false alarm rate 0.10 are given in Table 1.

Table 1: *The detection performance in changed environments under $P_f = 0.1$. The last row is tested with a different echo learning net.*

Case		Anet	Bnet	Cnet
baseline		81.7	83.0	79.6
NER	10 dB	96.6	96.8	95.3
	5 dB	91.9	92.8	90.1
	−5 dB	71.6	71.7	68.6
SNR	35 dB	82.2	82.6	79.4
	25 dB	79.2	80.2	77.6
	20 dB	72.8	72.6	71.6
RT	200 ms	80.9	79.7	78.9
	250 ms	77.4	76.3	75.2
	*250 ms	78.8	79.4	76.5

Seen from the results, the trained networks favor easier tasks as indicated by high NER and high SNR. Overall, the Bnet stays ahead in performance. By increasing the near-end speech power just 5 dB higher than the echo, double talk is more accurately detected with about 10 percents improvements. However, the networks are also badly affected by the low near-end speech power, which is also known as a big challenge in conventional methods. Similar trends are observed in relative to SNR, though the fluctuations are smaller. In the RT test, it is found that the networks are tolerable to small changes in the echo path. When the changes become bigger, the mismatch in the echo learning network and the classifier begins to add, which lowers the accuracies.

In the last test, we increase the variability in the data of class H_0, considering no-double-talk data is more easily obtained in reality. Several different echo paths are incorporated while the total amount of training data is unchanged. The three networks are retrained and the results are given in the last row. The multi-condition training setup alleviates the problem of data mismatch. Bnet here reports better results than the other two, which again supports the arguments in subsection 3.3.

5. Conclusion

In deep learning applications, a structured design is known beneficial to the final performance. In this paper, a separate echo learning net is combined with the DNN based classifier for the double talk detection problem. The echo learning net partially handles the nonlinearity in the system and collaborates with the classifier. The proposed networks outperform the conventional NCC method and they are shown robust to small changes in the environment. Future work may include network joint training and performance evaluation in AEC.

6. Acknowledgements

This work is supported by the China Scholarship Council (No. 201604910623).

7. References

[1] J. Benesty, C. Paleologu, T. Gänsler, and S. Ciochină, "Double-talk detection," in *A Perspective on Stereophonic Acoustic Echo Cancellation*. Springer, 2011, pp. 71–79.

[2] D. Duttweiler, "A twelve-channel digital echo canceler," *IEEE Transactions on Communications*, vol. 26, no. 5, pp. 647–653, 1978.

[3] C. Paleologu, J. Benesty, T. Gaensler, and S. Ciochină, "Class of double-talk detectors based on the holder inequality," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 425–428.

[4] M. Hamidia and A. Amrouche, "Double-talk detector based on speech feature extraction for acoustic echo cancellation," in *International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*. IEEE, 2014, pp. 393–397.

[5] H. Ye and B.-X. Wu, "A new double-talk detection algorithm based on the orthogonality theorem," *IEEE Transactions on Communications*, vol. 39, no. 11, pp. 1542–1545, 1991.

[6] C. Schuldt, F. Lindstrom, and I. Claesson, "A delay-based double-talk detector," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1725–1733, 2012.

[7] J. H. Cho, D. R. Morgan, and J. Benesty, "An objective technique for evaluating doubletalk detectors in acoustic echo cancelers," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 718–724, 1999.

[8] J. Benesty, D. R. Morgan, and J. H. Cho, "A new class of doubletalk detectors based on cross-correlation," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 168–172, 2000.

[9] M. A. Iqbal, J. W. Stokes, and S. L. Grant, "Normalized double-talk detection based on microphone and AEC error cross-correlation," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2007, pp. 360–363.

[10] T. Gänsler and J. Benesty, "The fast normalized cross-correlation double-talk detector," *Signal Processing*, vol. 86, no. 6, pp. 1124–1139, 2006.

[11] T. Gänsler, M. Hansson, C.-J. Ivarsson, and G. Salomonsson, "A double-talk detector based on coherence," *IEEE Transactions on Communications*, vol. 44, no. 11, pp. 1421–1427, 1996.

[12] I. J. Tashev, "Coherence based double talk detector with soft decision," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 165–168.

[13] G. Szwoch, A. Czyżewski, and M. Kulesza, "A low complexity double-talk detector based on the signal envelope," *Signal Processing*, vol. 88, no. 11, pp. 2856–2862, 2008.

[14] M. Z. Ikram, "Double-talk detection in acoustic echo cancellers using zero-crossings rate," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 1121–1125.

[15] H. K. Jung, N. S. Kim, and T. Kim, "A new double-talk detector using echo path estimation," *Speech communication*, vol. 45, no. 1, pp. 41–48, 2005.

[16] P. Åhgren, "Acoustic echo cancellation and doubletalk detection using estimated loudspeaker impulse responses," *IEEE Transactions on speech and audio processing*, vol. 13, no. 6, pp. 1231–1237, 2005.

[17] M. A. Iqbal, J. W. Stokes, J. C. Platt, A. C. Surendran, and S. L. Grant, "Doubletalk detection using real time recurrent learning," 2006.

[18] K.-H. Lee, J.-H. Chang, N. S. Kim, S. Kang, and Y. Kim, "Frequency-domain double-talk detection based on the gaussian mixture model," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 453–456, 2010.

[19] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.

[20] Q. Wang, J. Du, X. Bao, Z.-R. Wang, L.-R. Dai, and C.-H. Lee, "A universal vad based on jointly trained deep neural networks." in *INTERSPEECH*, 2015, pp. 2282–2286.

[21] C. M. Lee, J. W. Shin, and N. S. Kim, "DNN-based residual echo suppression." in *INTERSPEECH*, 2015, pp. 1775–1779.

[22] D. Comminiello, M. Scarpiniti, L. A. Azpicueta-Ruiz, J. Arenas-Garcia, and A. Uncini, "Functional link adaptive filters for nonlinear acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1502–1512, 2013.

[23] S. S. Haykin, *Adaptive filter theory*. Prentice Hall, 2002.