INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Countermeasures for Automatic Speaker Verification Replay Spoofing Attack : On Data Augmentation, Feature Representation, Classification and Fusion

Weicheng Cai[1], *Danwei Cai*[1,2], *Wenbo Liu*[1], *Gang Li* [3], *Ming Li* [1,2]

[1]SYSU-CMU Joint Institute of Engineering, School of Electronics and Information Technology,
Sun Yat-sen University, Guangzhou, China
[2]SYSU-CMU Shunde International Joint Research Institute, Guangdong, China
[3]Jiangsu Jinling Science and Technology Group Limited

liming46@mail.sysu.edu.cn

Abstract

The ongoing ASVspoof 2017 challenge aims to detect replay attacks for text dependent speaker verification. In this paper, we propose multiple replay spoofing countermeasure systems, with some of them boosting the CQCC-GMM baseline system after score level fusion. We investigate different steps in the system building pipeline, including data augmentation, feature representation, classification and fusion. First, in order to augment training data and simulate the unseen replay conditions, we converted the raw genuine training data into replay spoofing data with parametric sound reverberator and phase shifter. Second, we employed the original spectrogram rather than CQCC as input to explore the end-to-end feature representation learning methods. The spectrogram is randomly cropped into fixed size segments, and then fed into a deep residual netowrk (ResNet). Third, upon the CQCC features, we replaced the subsequent GMM classifier with deep neural networks including fully-connected deep neural network (FDNN) and Bidirectional Long Short Term Memory neural network (BLSTM). Experiments showed that data augmentation strategy can significantly improve the system performance. The final fused system achieves to 16.39 % EER on the test set of ASVspoof 2017 for the common task.

Index Terms: ASVspoof, replay attack, data augmentation, end-to-end, representation learning, ResNet

1. Introduction

Automatic speaker verification (ASV) refers to automatically accept or reject a claimed identity by his or her voice, and nowadays it is widely used in real-world biometric authentication applications [1, 2, 3]. However, a growing number of studies have confirmed the severe vulnerability of state-of-the-art ASV systems under a diverse range of intentional fraudulent attacks [4, 5, 6]. The initiative of the series ASVspoof challenge aims to promote the development of spoofing countermeasure studies [7]. The task in previous ASVspoof 2015 challenge was to discriminate genuine human speech from speech produced using text-to-speech and voice conversion attacks [8]. Arguably, however, replay attacks might be the most common spoofing technique to ASV especially for text dependent speaker verification, as it does not require the attackers to have any speech technology knowledge and can be mounted with greater ease using common consumer devices [9, 10].

The ongoing ASVspoof 2017 challenge is to assess audio replay spoof attack detection 'in the wild' [11]. The task is to determine whether a given speech audio is genuine human voice or replayed recording. The challenge is focused on the development of generalized and robust spoofing attack detectors with the capability of detecting various of replay attacks with both known and unknown conditions [12].

Recently, a new constant Q cepstral coefficient(CQCC) feature based on the constant Q transform(CQT), which is a perceptually-inspired time-frequency analysis tool popular in the music study, was proposed to detect various kinds of spoofing attacks [13, 14]. It is shown in [14, 15] that CQCC outperforms many previously reported features by a significant margin against both known and unknown attacks. It is further studied that there is more gain that could be achieved by designing effective feature representations rather than investigating more advanced or complex classifiers with common features. Concretely, the standard Gaussian Mixture Model (GMM) trained with maximum likelihood criterion has been shown to yield among the best performances, compared with various kind of generative and discriminative methods including GMM-UBM [1], GLDS-SVM [16], GMM-SVM [16], i-vectors [17], etc., given the short duration audio inputs.

Based on the state-of-the-art CQCC-GMM method, we have investigated different steps in the countermeasure system building pipeline, including data augmentation, feature representation, classification and fusion. The motivation behind is that introducing multiple diverse, competitive, and complementary methods could potentially boost the baseline performance significantly after score level fusion.

As the first contribution of this paper, we proved the effectiveness of artificial data augmentation strategy. We generated a set of "spoof-liked" audio samples through different parametric reverberators and phase shifters to simulate the real world replay attack channel characteristics. It is shown that the GMM trained with pooled 'real' spoofing data and 'simulated' spoofing data can capture more pattern of unknown conditions. The second contribution is that we introduced an end-to-end representation learning framework rather than following the conventional handcrafted feature based methods. We directly fed the original audio spectrogram into a deep ResNet, thus the feature descriptor and classifier can be learned in an aggregated end-to-end manner. To the best of our knowledge, there are only few existing end-to-end spoof countermeasure systems and the proposed ResNet framework presents a potentially new direction for automatic feature learning especially with large amount of

This research was funded in part by the National Natural Science Foundation of China (61401524), Natural Science Foundation of Guangdong Province (2014A030313123), Natural ScienceFoundation of Guangzhou City(201707010363), the Fundamental Research Fundsfor the Central Universities(15lgjc12) National Key Research andDevelopment Program (2016YFC0103905) and IBM Faculty Award.

http://dx.doi.org/10.21437/Interspeech.2017-906

training data which might be available in future. Last but not least, we came to the same conclusion as [14, 18] that although GMM back-end can't achieve the best performance on development dataset, it has strong capability in anti-overfitting and always superior on the test set, which contains a number of data from various kinds of unknown conditions.

2. Methods

2.1. Data augmentation

In order to simulate the probably unseen replay condition, we converted the raw genuine training data to simulated replay spoofing data by some parametric sound reverberator and phase shifter for data augmentation.

Compared to the genuine human voice, the replayed speech generally has some special acoustic characteristics related to the loudspeaker, the room reverberation and the microphone. Since our training data can not cover all the replay conditions, our model should have good generability for unseen conditions.

In general, when people directly speak to the microphone, the strong air flow coming out of the mouth makes the collected speech contain high percentage of directly arrived sound with less reverberation [19]. However, the common loudspeakers used in the replaying attacks do not have the acoustic vocal effect of human talking head. The replayed speech will inevitably introduce more reverberation unless the line-out to line-in recording channel is used. This motivates us to use a parametric reverberator to artificially simulate ”spoof-liked” speech from the genuine speech.

Moreover, since the original speech signal went through a complex replay pipeline, there might be some distortions brought by the imperfect playing and recording devices or environments. Therefore, we also adopt a phaser [20] to simulate some distorted speech as a part of the simulated spoofing data.

We use the Adobe Audition CC software with the default parameter setup to simulate the effects of reverberator and phaser. For each genuine speech, we generate two simulated spoofing data using the parameters shown in Table 1. By adding these simulated replay spoofing data in the training set, the system becomes less over-fitting.

Table 1: *Parameters of studio reverberation effect and phaser effect in the Adobe Audition software*

studio reverberation		phaser	
Room Size	100	Intensity	100%
Decay	2000ms	Depth	72%
High Frequency Cut	897 Hz	Mod Rate	2.43 Hz
Low Frequency Cut	385 Hz	Upper Freq	54 Hz
Damping	80%	Feedback	64 %
Diffusion	20%	Output Gain	-3.3dB

2.2. Feature representation

2.2.1. Handcrafted CQCC feature

The so-called CQCC feature is obtained by perceptually-aware CQT coupled with traditional cepstral analysis. The extraction framework is shown in Fig. 1, more details of CQCC can be found in [14].

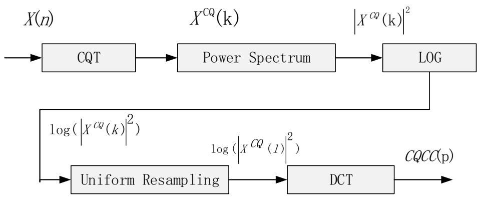
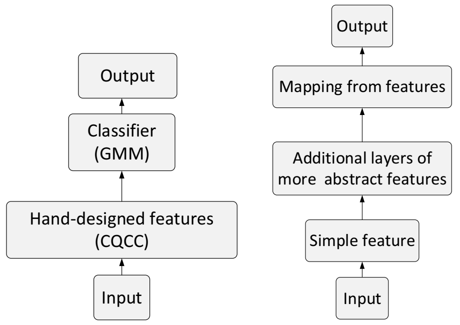


Figure 1: *Block diagram of CQCC feature extraction*

2.2.2. Representation learning upon ResNet

As shown in Fig. 2, traditional machine learning methods might have to build classifiers on hand-designed features, which requires extensive domain knowledge from human experts. Representation learning, on the opposite, tries to represent the signal as a nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones [21].



(a) classic machine learning (b) end-to-end representation learning

Figure 2: *Flowcharts of handcrafted feature modeling and end-to-end representation learning*

ResNet have emerged as a family of very deep architectures showing competitive accuracy and nice convergence behaviors in many computer vision tasks such as object recognition, face identification, emotion recognition [22, 23]. They are neural networks in which each layer consists of a residual module f_i and a skip connection bypassing f_i. Since layers in ResNet can compromise multiple convolutional layers, they are referred to as residual block, which is shown in Fig. 3.

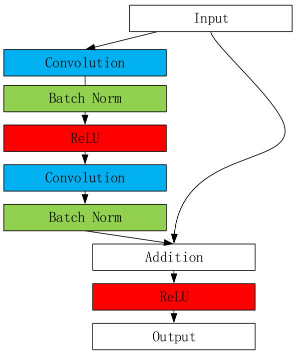


Figure 3: *An example of typical block*

With x as input, the output of the ith block is recursively defined as

$$y_i \equiv f_i(x) + x \tag{1}$$

where $f_i(x)$ is a sequence of operations convolutions, batch normalization, and rectified linear units(RELU). In the most recent formulation of ResNetf, $f_i(x)$ is defined by

$$f_i(x) \equiv (B(W_i \cdot \sigma(B(W_i^{'} \cdot x)))) \tag{2}$$

where W_i and $W_i^{'}$ and are weight matrices,· denotes convolution, $B(x)$ is batch normalization and $\sigma(x) \equiv max(x, 0)$.

Table 2: *ResNet Configuration*

layer_name	output size	34-layer
conv1	112x112	7x7, 64, stride 2
conv2_x	56x56	3x3 max pool, stride 2 $\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 3$
conv3_x	28x28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 4$
conv4_x	14x14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 6$
conv5_x	7x7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 3$
	1x1	average pool, 2-d fc, softmax

As for ResNet, we randomly cropped multiple fix sized images (224x224) from the Short-time Fourier transform(STFT) based spectrogram and then fed them into a standard 34 layer ResNet as shown in Table 2.

2.3. Classification

After obtaining the feature representation of each utterance, we need to train a robust classifier to detect the replay recordings. In this section, we investigate different classification methods based on CQCC feature.

2.3.1. GMM

GMM is a stochastic generative model, and is widely used to model the probability distribution of audio features.

In the test phase, given the models, $\lambda_{genuine}$ and λ_{spoof}, and the feature vectors of the test utterance $\boldsymbol{Y} = [y_1, ...y_T]$, the detection score is computed as follows [18]:

$$\Lambda(Y) = \Gamma(\boldsymbol{Y}|\lambda_{genuine})) - \Gamma(\boldsymbol{Y}|\lambda_{genuine})). \tag{3}$$

where $\Gamma(\boldsymbol{Y}|\lambda) = (1/T)\sum_{t=1}^{T} \log p(\boldsymbol{y}_t|\lambda)$ is the average log-likelihood of $\boldsymbol{Y}$ given GMM model λ. $\lambda_{genuine}$ and $\lambda_{spoof}$are the GMMs for genuine and spoofed classes, respectively.

For the baseline system, we followed the matlab implementation of CQCC extraction together with GMM classifier provided by [14]. Every audio sample is converted to a 90 dimensional CQCC feature sequence. Then, two 512-component GMMs are trained on the genuine and spoofed speech utterances, respectively. The score for a given test utterance is computed as log-likelihood ratio between these two GMM models.

2.3.2. FDNN

Although GMM can model the probability distribution of given features, as a kind of generative model,it may not be optimum in terms of discrimination [24]. Besides, the feature vectors for GMM are assumed to be independent and identically distributed which might not be true in our case. As a result, it can't exploit the correlated information embedded in the context. With this consideration, we keep the CQCC as input feature, replacing GMM with FDNN, as shown in Fig. 4. Each feature vector is concatenated with its partial left 4 context window and right 4 context window feature vectors. These feature vectors are then flattened into a single 810 dimension vector as input of the FDNN. The output layer has 1 units, and the binary cross-entropy loss is adopted. Similar to GMM system, the ultimate score was computed from the mean pooling of the frame level posterior probabilities.

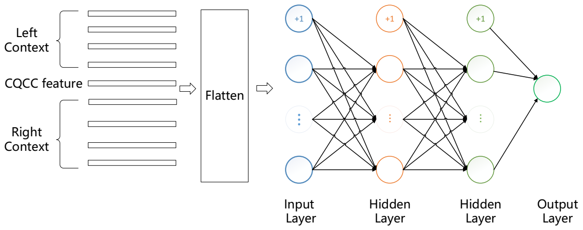


Figure 4: *FDNN arthitecture for input CQCC feature*

2.3.3. BLSTM

The third classifier investigated is BLSTM. Given an input sequence $\boldsymbol{x} = [x_1, ..., x_T]$ and the hidden vector $\boldsymbol{h} = [h_1, ..., h_T]$, for a standard recurrent neural networks(RNNs), the output vector $\boldsymbol{y} = [y_1, ..., y_T]$ can be computed from $t = 1$ to T according to the following iterative equations:

$$h_t = H(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{4}$$

$$y_t = W_{ht}h_t + b_y \tag{5}$$

where H is the activation function of hidden layer, W is the weight matrix, and b is the bias vectors.

Bidirectional RNNs(BRNNs) were proposed to make full use of the context of feature sequences in both forward and backward directions [25]. Furthermore, an LSTM structure consists of memory blocks was proposed to learn the long term dependencies [26, 27]. Every block contains self-connected memory cells and three adaptive and multiplicative gate units i.e. input, output,and forget gates. These gates can respectively provide write, read, reset operations for the cells. After combining the advantages of BRNN and LSTM, BLSTM [28], designed as Fig. 5, can deal with long-range context in both preceding and succeeding directions.

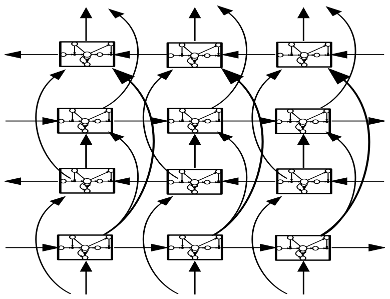


Figure 5: *Typical BLSTM structure*

Since BLSTM is just considered to build up high level representation of input features, additional fully-connected layer is needed to map it into binary categorical output. We chose a large window contained left 20 frames and right 20 frames context,compare with FDNN. Therefore, a 41×90 sequential feature is derived to feed into the BLSTM network.

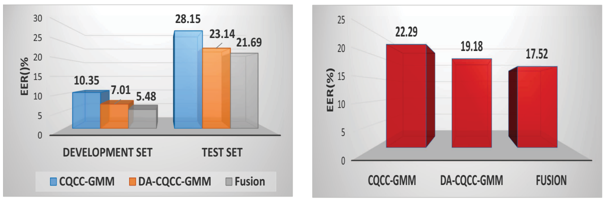
3. Experiments

3.1. Data protocol

We followed the original data participation protocol provided by the ASVspoof 2017 challenge organizers. The training dataset contains 3016 utterances including 1508 genuine utterances and 1508 spoofing utterances from 10 speakers. The development dataset contains 1710 utterances including 760 genuine utterances and 950 spoofing utterances from 8 speakers. The test dataset with unknown genuine/spoof label contains totally 14220 audio samples. Most of our implemented systems are trained in two versions, one is trained by only train data, the other by using pooled train and development data.

3.2. Results on data augmentation

The results in Fig. 6 reveal that data augmented CQCC-GMM(DA-CQCC-GMM) outperforms the CQCC-GMM baseline by approximate 30% relatively on development set. On test set, 18% relative performance improvement is gained when trained without development data, meanwhile 14% relative performance gained in pooling development data condition. After score level fusion, the system achieves to 17.52% EER, which is a relatively 23% performance improvement.



(a) system performance trained by only training data

(b) system performance trained by pooled train+development data

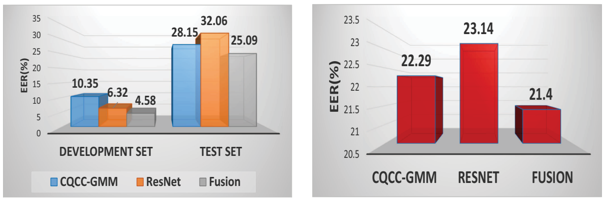
Figure 6: *Results on data augmentation*

3.3. Results on end-to-end representation learning

Here shows the results on end-to-end representation learning in Fig. 7. It reveals that end-to-end ResNet method outperform CQCC-GMM significantly on development set. For test set, in the contrast, CQCC-GMM slightly superior to ResNet, and after score level fusion, system performance is boosted.

3.4. Results on classifiers

Experiment results in Table 3 show that although deep learning methods like FDNN and LSTM can achieve significantly superior performance compared with GMM on development set, they decline sharply on test set. The BLSTM got 40.08% EER and the FDNN got nearly almost all the posteriors to zero (thereby we didn't submit to the challenge organizer the FDNN system results on test set), with both of them reveal severe overfitting.



(a) system performance trained by only training data

(b) system performance trained by pooled train+development data

Figure 7: *Results on end-to-end representation*

Table 3: *System performance by different classifier*

Classifier	EER on Devel. Set (%)	EER on Test Set (%)
GMM	10.35	**28.15**
FDNN	6.41	-
BLSTM	**5.82**	40.08

3.5. System fusion results

Finally, as presented in Table 4, after score level fusion on the results of CQCC-GMM, DA-CQCC-GMM and ResNet, the proposed method achieve 16.39% EER performance, which outperforms baseline by 26% relatively .

Table 4: *Final system fusion results*

System	Devel. Set (%)	Test Set (%)
CQCC-GMM(baseline)	10.35	22.29
DA-CQCC-GMM	7.01	19.18
ResNet	6.32	23.14
Score level fusion	**3.52**	**16.39**

4. Conclusions and future works

This paper investigates different steps in the ASV spoof countermeasure system building pipeline, including data augmentation, feature representation, classification and fusion. It shows the effectiveness of simulating the unknown 'spoof-liked' data, therefore drives us to pursuit higher generalization ability on small limited data through various kinds of data augmentation strategy. Besides, the comparable performance produced by ResNet reveals a possible good potential of end-to-end representation learning, which requires little human experts' domain knowledge.

In the future, there remains much to be done: (1) The data augmentation strategy in this paper is done manually, and quite rely on human knowledge. It is possible to seek a data driven generative adversarial models to automatically learn the pattern of 'spoof-liked' data; (2) In the experiments, we only use the STFT based spectrogram as the input of ResNet, it might not be optimal and we can try to investigate some perceptually-aware spectrogram like CQT spectrogram, Gammatone spectrogram, etc.; (3) Although FDNN/BLSTM is inferior to GMM in the experiment, the significant improvement on development dataset motivates us to investigate its strong representation ability on larger scale datasets.

5. References

[1] D. A Reynolds and R. C Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech & Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[3] J. H L Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[4] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Proc. Interspeech 2013*, 2013.

[5] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[6] S. K Ergunay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *IEEE International Conference on Biometrics Theory, Applications and Systems*, 2015, pp. 1–6.

[7] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilci, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "Asvspoof: the automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. PP, no. 99, pp. 1–1, 2017.

[8] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilc, Sahidullah M., and S. Aleksandr, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech 2015*, 2015.

[9] Z. Wu and H. Li, "On the study of replay and voice conversion attacks to text-dependent speaker verification," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5311–5327, 2016.

[10] A. Janicki, F. Alegre, and N. Evans, "An assessment of automatic speaker verification vulnerabilities to replay spoofing attacks," *Security and Communication Networks*, vol. 9, no. 15, pp. 3030–3044, 2016.

[11] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamki, D. Thomsen, A. Sarkar, Z. Tan, H. Delgado, and M. Todisco, "Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *Proc. ICCASP 2017*, 2017.

[12] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech 2017*, 2017.

[13] J. C. Brown, "Calculation of a constant q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[14] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," in *Odyssey 2016 - The Speaker and Language Recognition Workshop*, 2016.

[15] M. Todisco, H. Delgado, and Nicholas Evans, "Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech and Language*, 2017.

[16] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[17] N. Dehak, P. J Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[18] C. Hanili, T. Kinnunen, M. Sahidullah, and A. Sizov, "Classifiers for synthetic speech detection: A comparison," in *Proc. Interspeech 2015*, 2015.

[19] J. Mcdonough and M. Wolfel, "Distant speech recognition: Bridging the gaps," in *Hands-Free Speech Communication and Microphone Arrays*, 2008, pp. 108–114.

[20] S. Wardle, "A hilbert-transformer frequency shifter for audio," *First Workshop on Digital Audio Effects Dafx*, 1970.

[21] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR 2016*, 2016, pp. 770–778.

[23] A. Veit, M. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," in *Proc. Advances in Neural Information Processing System*, 2016.

[24] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T. N. Sainath, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[25] M. Schuster and K. K Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[26] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm. neural computation 12(10): 2451-2471," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[27] A. Graves, *Long Short-Term Memory*, Springer Berlin Heidelberg, 2012.

[28] A. Graves, N. Jaitly, and A. R. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *Proc. ASRU*, 2013, pp. 273–278.