INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual neTworks

Cheng-I Lai, Nanxin Chen, Jesús Villalba, Najim Dehak

Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
{clai24,bobchennan,jvillal7,ndehak3}@jhu.edu

Abstract

We present JHU's system submission to the ASVspoof 2019 Challenge: Anti-Spoofing with Squeeze-Excitation and Residual neTworks (ASSERT). Anti-spoofing has gathered more and more attention since the inauguration of the ASVspoof Challenges, and ASVspoof 2019 dedicates to address attacks from all three major types: text-to-speech, voice conversion, and replay. Built upon previous research work on Deep Neural Network (DNN), ASSERT is a pipeline for DNN-based approach to anti-spoofing. ASSERT has four components: feature engineering, DNN models, network optimization and system combination, where the DNN models are variants of squeeze-excitation and residual networks. We conducted an ablation study of the effectiveness of each component on the ASVspoof 2019 corpus, and experimental results showed that ASSERT obtained more than 93% and 17% relative improvements over the baseline systems in the two sub-challenges in ASVspooof 2019, ranking ASSERT one of the top performing systems. Code and pre-trained models are made publicly available[1].

Index Terms: ASVspoof, Anti-Spoofing, Speaker Verification

1. Introduction

Automatic Speaker Verification (ASV) has become an increasingly attractive option for biometric authentication. Past research has shown that ASV systems are subject to malicious attacks: the presentation attacks. Presentation attacks, or spoofing attacks, refer to attempts of bypassing ASV systems by mimicking the voice characteristics of the target speaker. Spoofing attacks have four widely-recognized specifications: impersonation, replay, text-to-speech (TTS) and voice conversion (VC). To defend against these attacks, a standalone anti-spoofing system is developed in parallel to the ASV system [1]. Recent efforts on anti-spoofing developments mainly originated from the Biennial ASVspoof Challenges [2, 3, 4].

Previous ASVspoof Challenges focused on promoting awareness and fostering solutions to spoofing attacks generated from TTS, VC and replay [2, 3, 4]. ASVspoof 2019 aims to address all previous attacks and further extended previous editions of ASVspoof in three aspects:

- Update attacks with TTS and VC with state-of-the-art technologies, especially those based on neural networks.
- Create a more controlled setup for replay attacks, covering acoustic and microphone conditions and predefined replay device qualities.
- Adopt an evaluation metric to assess impacts of standalone anti-spoofing systems to a fixed ASV system.

ASVspoof 2019 Challenge is composed of two sub-challenges: Physical Access (PA) and Logical Access (LA). LA considers spoofing attacks generated with TTS and VC, and PA refers to spoofing attacks from replay.

[1] https://github.com/jefflai108/ASSERT

Research work on anti-spoofing can be divided into one of the three categories: Feature Learning [5, 6, 7, 8, 9, 10, 11, 12, 13, 14], Statistical Modeling [4, 15, 16, 17], and Deep Neural Network (DNN) [18, 19, 20, 21, 22, 23, 24, 25]. Having witnessed the successes of DNNs in ASVspoof 2017, we decided to explore and extend several DNN-based systems for the ASVspoof 2019 Challenge. Our objective is to identify and design core components of a working pipeline for DNN-based approach to anti-spoofing. These components, feature engineering, DNN models, network optimization, and system fusion, make up our anti-spoofing system, which we term Anti-Spoofing with Squeeze-Excitation and Residual neTworks, or ASSERT. The main contribution of this paper is two-fold:

1. We conducted experiments on the effectiveness of several DNN models in detecting spoofing attacks generated from audio replay, TTS and VC. The DNN models are based on variants of Squeeze-Excitation Network (SENet) [26] and ResNet [27]. To our knowledge, we were the first to introduce SENet and ResNet with statistical pooling to address anti-spoofing, and we also extended our previous work in [20] such that the DNNs are deeper but faster-trained.

2. We presented an ablation study, from feature engineering, network optimization, to fusion schemes for training DNN models for anti-spoofing. We believe these collective strategies are vital for the performance of DNNs. In addition, we compared ASSERT with our implementation of i-vectors baselines [28]. Results on the ASVspoof 2019 corpus demonstrated that ASSERT achieved significant performances over the baseline systems, with more than 93% and 17% relative improvements on PA and LA respectively. Our fusion system was ranked 3^{rd} in the PA sub-challenge, and 14^{th} in the LA sub-challenge.

The outline of the paper is organized as follows. Section 2 details ASSERT, from the feature engineering approaches, proposed DNN models, to the optimization and fusion schemes. Section 3 compares the results of ASSERT with the baseline systems on the ASVspoof 2019 corpus. We ended the paper with some concluding remarks in Section 4.

2. ASSERT

This section presents an overview of each component of ASSERT: input feature representations to DNN models, the DNN models and their parameters, along with the network optimization and fusion schemes. The feature preparation is either a unified feature map or the whole utterance. Both approaches are based on some low-level acoustic features. The DNN models are variants of squeeze-excitation and residual networks: SENet34, SENet50, Mean-Std ResNet, Dilated ResNet, and Attentive-Filtering Network.

http://dx.doi.org/10.21437/Interspeech.2019-1794

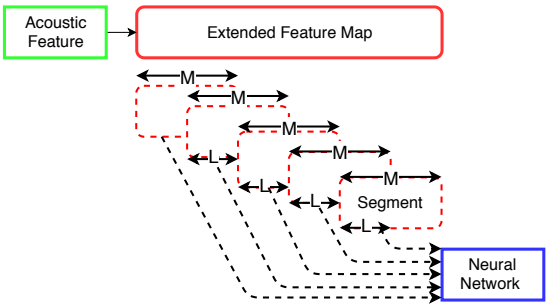

Figure 1: *Illustration of Unified Feature Map approach. Low-level acoustics feature are first extracted, and the utterance is repeated to form a unified feature map. Then, the feature map is broken down into segments with length M frames and overlap L frames, before inputting into the DNN models.*

2.1. Feature Engineering

Acoustic Features: We extracted two different acoustic features: constant Q cepstral coefficients (CQCC) [5] and log power magnitude spectra (logspec). Following [4], we extracted 30 dimension CQCC feature, including the 0'th order cepstral coefficient and without CMVN. The dimension of logspec is 257. For both CQCC and logspec, we did not apply voice activity detection nor any normalization to the acoustic features, as we empirically found doing so yield better results.

Unified Feature Map[2]**:** We followed previous work [20] and created a unified feature map as input to the DNN models. Since the lengths of evaluation utterances were not known beforehand, we first extended all utterances to multiple of M frames. Then, the extended feature map was broken down into segments of length M frames. The segments can have L frames overlap.

For the 2019 ASVspoof Challenge, M is set to 400, and L is set to either 0 or 200. Figure 1 is an illustration of this feature engineering approach. There may be multiple segments per utterance. We simply averaged the DNN outputs over all segments for each utterance.

Whole Utterance: In addition to the Unified Feature Map, we considered another feature engineering approach by training models with the whole utterance (variable length input). For each minibatch during training, utterances are zero-padded to match the length of the longest utterance. Padding frames are subsequently removed in the pooling layer of the DNNs.

2.2. DNN model

Squeeze-Excitation Network: Given recent achievements in spoofing countermeasures from different DNN architectures [19, 20], we explored an extension of ResNet, Squeeze-Excitation Network (SENet), for ASVspoof 2019. SENet has attained impressive image classification results, where a channel-wise transform is appended to existing DNN building blocks, such as the Residual unit [26]. We implemented two variants of SENets: SEnet34 with ResNet34 backbone, and SEnet50 with ResNet50 backbone. Table 1 contains the model parameters of SEnet34 and SEnet50. SEnet34 and SEnet50 were trained with unified feature maps of logspec while each minibatch contains 64 feature maps.

Mean-Std ResNet: Recent work in speaker recognition [29, 30] has demonstrated that ResNet [27] with pooling achieves comparable results as x-vectors [31]. Therefore, we introduced ResNet with pooling for anti-spoofing. Specifically, we employed Mean-Std ResNet, where mean and standard deviation are estimated over timesteps to represent the whole utterance [31] after frame-level features are extracted from a ResNet34. Table 1 contains the model parameters of a Mean-Std ResNet. Since the pooling layer accounts for variable length input, we train Mean-Std ResNet with the whole utterance. Both CQCC and logpsec were used, while each minibatch contains 64 and 32 full utterances, respectively.

[2]This is not practical for long utterances, but spoofed speech are mostly recorded in short duration (less than 10 seconds).

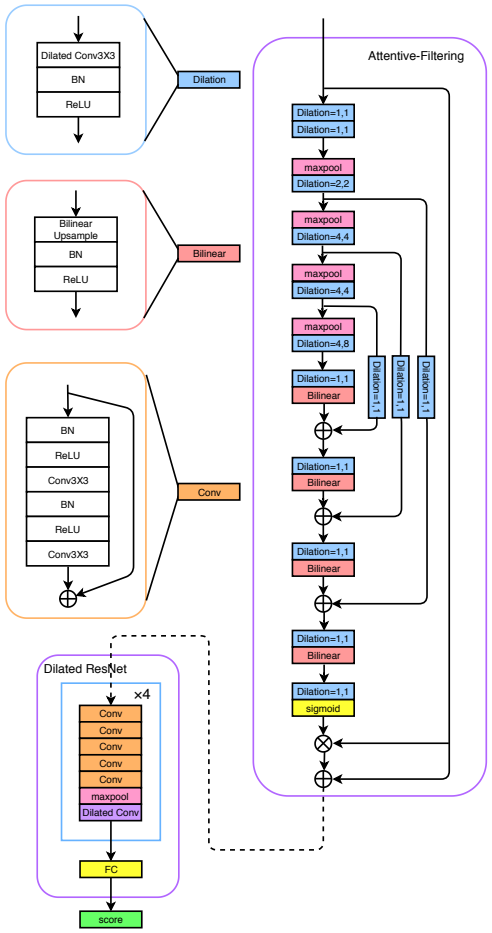

Figure 2: *(Bottom Left) Dilated ResNet is consisted of four blocks followed by a fully-connected layer. Each block has five residual units, a max-pooling layer, and a dilated convolution layer. (Right) Attentive-Filtering applies an attention-based masking prior to a Dilated ResNet. Input feature goes through four downsampling and four upsampling units. Skip connection is used throughout. Dilation indicates the dilation rate of each convolution layer. Bilinear indicates bilinear upsampling.*

Dilated ResNet: Following previous work [20], we applied Dilated ResNet to ASVspoof 2019. Different from Mean-Std ResNet, Dilated ResNet contains a dilated convolution layer in each residual block [32]. We also extended the original dilated residual block to multiple residual units. Figure 2 is a sketch of the Dilated ResNet, and Table 1 contains its model parameters. Contrary to Mean-Std ResNet, Dilated ResNet does not have any pooling layer and thus only accepts fixed-size input. We trained Dilated ResNet with the same condition as the SENets.

Table 3: *Ablation study of single system results on ASVspoof 2019. Due to space constraint, for each DNN model, we merely included* ***top two*** *performing systems in the table. Under Training Objective column, MCE stands for multi-class cross entropy, BCE stands for binary cross entropy, and acc/EER stands for the model selection criterion after each training epoch.*

Model	Acoustic Feature	Feature Engineering	Training Objective	Model Params.	PA development		LA development	
					t-DCF$^{min}_{norm}$	EER (%)	t-DCF$^{min}_{norm}$	EER (%)
CQCC-GMM	CQCC + Δ + $\Delta\Delta$	N/A	EM	138k	0.195	9.87	0.012	0.43
LFCC-GMM	LFCC + Δ + $\Delta\Delta$	N/A	EM	92k	0.255	11.96	0.066	2.71
100-i-vectors	CQCC + Δ + $\Delta\Delta$	N/A	EM	593k	0.306	12.37	0.155	5.18
200-i-vectors	CQCC + Δ + $\Delta\Delta$	N/A	EM	2339k	0.322	12.52	0.121	4.12
SENet34	logspec	unifed, L=200	BCE + acc.	1344k	**0.015**	**0.575**	**0**	**0**
	logspec	unifed, L=200	BCE + EER	1344k	0.017	0.686	0	0
SENet50	logspec	unifed, L=200	MCE + EER	1095k	0.021	0.799	0	0
	logspec	unifed, L=200	BCE + EER	1093k	0.017	0.631	0	0
Mean-Std ResNet	logspec	whole	BCE + acc.	1389k	0.022	0.832	0	0
	CQCC	whole	MCE + acc.	1390k	0.041	1.429	0.001	0.040
Dilated ResNet	logspec	unifed, L=200	MCE + EER	593k	0.029	1.072	0	0
	logspec	unifed, L=200	BCE + EER	592k	0.024	0.780	0	0
Attentive-Filteirng Net	logspec	unifed, L=200	MCE + EER	600k	0.027	1.057	0	0
	logspec	unifed, L=200	BCE + acc.	599k	0.021	0.740	0	0

verification system, provided by the organizers in the case of ASVspoof 2019 Challenge.

Implementation: We used training partition to train our DNN models. Development partition was used for model selection during validation and system combination. We did not use any external data or data augmentation technique for development. CQCC-GMM and LFCC-GMM were adopted directly from the MATLAB script[3]. Acoustic features and i-vectors were extracted with Kaldi [37]. DNNs were implemented in PyTorch.

3.3. Ablation Study of Single Systems

Table 3 compares ASSERT with the baseline systems in spoofing countermeasure on the *dev* partition. The first observation is that the i-vectors baseline performed worse than GMMs surprisingly, which is contrary to prior work on the ASVspoof 2017 corpus [4, 15]. ASSERT attains substantial improvements from the baseline GMM and i-vectors systems on both PA and LA. In general, for training the proposed DNN models, logspec outperforms CQCC, and unified feature map with overlap outperforms without overlap and whole utterance. On the other hand, there are mixed results on using multi-task or binary training objective, and on model selection with *dev* EER or *dev* classification accuracy. We empirically found that the best single system is based on SENet34 trained with unified feature map with overlap of logspec and binary cross-entropy loss with *dev* accuracy model selection. The system obtains 92% and 94% relative improvements over CQCC-GMM on *dev* t-DCF and EER for PA, and 100% relative improvements for LA.

3.4. Evaluation Results

Table 4 is the summary of our primary and single system submission to the ASVspoof 2019 Challenge. The single system is based on SENet34 (logspec), and the primary system is a system combination of five single systems based on SENet34 (logspec), Mean-Std ResNet (CQCC, logspec), SENet50 (logspec) and Dilated ResNet (logspec). Systems are trained separately for PA and LA. We can observe that ASSERT generalizes well across *dev* and *eval* for PA, nevertheless, it overfits on *dev* for LA. Our primary system further gains 93% and 95% relative improvements over CQCC-GMM on *eval* t-DCF and EER for PA, and 27% and 17% relative improvements over LFCC-GMM on *eval* t-DCF and EER for LA.

[3]http://www.asvspoof.org

Table 4: *Primary, single and baseline systems for ASVspoof 2019. Single system is based on SENet34; primary system is a fusion of five systems based on: SENet34, Mean-Std ResNet (CQCC, logspec), SENet50 and Dilated ResNet.*

System	Development		Evaluation	
	t-DCF$^{min}_{norm}$	EER	t-DCF$^{min}_{norm}$	EER
PA-single	0.015	0.575	0.036	1.29
PA-primary	**0.003**	**0.129**	**0.016**	**0.59**
PA-baseline	0.195	9.87	0.245	11.04
LA-single	0	0	0.216	11.75
LA-primary	**0**	**0**	**0.155**	**6.70**
LA-baseline	0.066	2.71	0.212	8.09

4. Conclusions

We introduced ASSERT – several variants of squeeze-excitation and residual networks, optimization and fusion schemes, along with feature engineering approaches – for anti-spoofing. Our fusion system attained considerable improvement over baseline systems on the ASVspoof 2019 corpus. We believe this paper serves as a preliminary work on a more comprehensive study on DNN based countermeasures for speech spoofing attacks, while meta-data analysis and model refinements on LA should be further investigated.

Acknowledgments The authors thank ASVspoof 2019 committee for designing the corpus and organizing the challenge.

5. References

[1] 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.

[2] N. W. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification." in *Interspeech*, 2013, pp. 925–929.

[3] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[4] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. Lee, and J. Yamagishi, "Asvspoof 2017 version 2.0: meta-data analysis and baseline enhancements," in *Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018.

[5] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," in *Speaker Odyssey Workshop, Bilbao, Spain*, vol. 25, 2016, pp. 249–252.

[6] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," 2015.

[7] M. J. Alam, G. Bhattacharya, and P. Kenny, "Boosting the performance of spoofing detection systems on replay attacks using q-logarithm domain feature normalization," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 393–398.

[8] M. Saranya and H. A. Murthy, "Decision-level feature switching as a paradigm for replay attack detection."

[9] G. Suthokumar, V. Sethu, C. Wijenayake, and E. Ambikairajah, "Modulation dynamic features for the detection of replay attacks," in *Proc. Interspeech*, 2018, pp. 691–695.

[10] H. Sailor, M. Kamble, and H. Patil, "Auditory filterbank learning for temporal modulation features in replay spoof speech detection," in *Proc. Interspeech*, 2018, pp. 666–670.

[11] L.-W. Chen, W. Guo, and L.-R. Dai, "Speaker verification against synthetic speech," in *2010 7th International Symposium on Chinese Spoken Language Processing*. IEEE, 2010, pp. 309–312.

[12] P. L. D. Leon, B. Stewart, and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[13] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[14] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7234–7238.

[15] M. Adiban, H. Sameti, N. Maghsoodi, and S. Shahsavari, "Sut system description for anti-spoofing 2017 challenge," in *Proceedings of the 29th Conference on Computational Linguistics and Speech Processing (ROCLING 2017)*, 2017, pp. 264–275.

[16] Z. Wu, T. Kinnunen, E. S. Chng, H. Li, and E. Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2012, pp. 1–5.

[17] E. Khoury, T. Kinnunen, A. Sizov, Z. Wu, and S. Marcel, "Introducing i-vectors for joint anti-spoofing and speaker verification," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[18] N. Chen, Y. Qian, H. Dinkel, B. Chen, and K. Yu, "Robust deep feature for spoofing detectionthe sjtu system for asvspoof 2015 challenge," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[19] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks." in *Interspeech*, 2017, pp. 82–86.

[20] C.-I. Lai, A. Abad, K. Richmond, J. Yamagishi, N. Dehak, and S. King, "Attentive filtering networks for audio replay attack detection," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[21] G. Valenti, H. Delgado, M. Todisco, N. Evans, and L. Pilati, "An end-to-end spoofing countermeasure for automatic speaker verification using evolving recurrent neural networks," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 288–295.

[22] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "Resnet and model fusion for automatic spoofing detection." in *INTERSPEECH*, 2017, pp. 102–106.

[23] B. Chettri, S. Mishra, B. L. Sturm, and E. Benetos, "A study on convolutional neural network based end-to-end replay anti-spoofing," *arXiv preprint arXiv:1805.09164*, 2018.

[24] H.-J. Shim, J.-W. Jung, H.-S. Heo, S.-H. Yoon, and H.-J. Yu, "Replay spoofing detection system for automatic speaker verification using multi-task learning of noise classes," in *2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. IEEE, 2018, pp. 172–176.

[25] Y. Qian, N. Chen, and K. Yu, "Deep features for automatic spoofing detection," *Speech Communication*, vol. 85, pp. 43–52, 2016.

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

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

[29] W. Cai, Z. Cai, X. Zhang, X. Wang, and M. Li, "A novel learnable dictionary encoding layer for end-to-end language identification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5189–5193.

[30] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, F. Richardson, S. Shon, F. Grondin *et al.*, "The jhu-mit system description for nist sre18."

[31] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[32] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[34] N. Brümmer and E. De Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," *arXiv preprint arXiv:1304.2865*, 2013.

[35] D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matějka, "Language recognition in ivectors space," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[36] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-dcf: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," *arXiv preprint arXiv:1804.09618*, 2018.

[37] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.