INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Blind speech signal quality estimation for speaker verification systems

Galina Lavrentyeva[1], *Marina Volkova*[1,2], *Anastasia Avdeeva*[2], *Sergey Novoselov*[1,2], *Artem Gorlanov*[2], *Tseren Andzukaev*[2], *Artem Ivanov*[2], *Aleksandr Kozlov*[2]

[1]ITMO University, St. Petersburg, Russia
[2]STC-innovations Ltd., St. Petersburg, Russia

{lavrentyeva, volkova, avdeeva-a, novoselov,gorlanov,andzhukaev,ivanov-ar, kozlov-a}@speechpro.com

Abstract

The problem of system performance degradation in mismatched acoustic conditions has been widely acknowledged in the community and is common for different fields. The present state-of-the-art deep speaker embedding models are domain-sensitive. The main idea of the current research is to develop a single method for automatic signal quality estimation, which allows to evaluate short-term signal characteristics.

This paper presents a neural network based approach for blind speech signal quality estimation in terms of signal-to-noise ratio (SNR) and reverberation time (RT60), which is able to classify the type of underlying additive noise. Additionally, current research revealed the need for an accurate voice activity detector that performs well in both clean and noisy unseen environments. Therefore a novel neural network VAD based on U-net architecture is presented.The proposed algorithms allow to perform the analysis of NIST, SITW, Voices datasets commonly used for objective comparison of speaker verification systems from the new point of view and consider effective calibration steps to improve speaker recognition quality on them.

Index Terms: Blind speech quality estimation, SNR, RT60, VAD, speaker recognition

1. Introduction

The quality of speaker recognition (SR) systems in real life vary according to the acoustic conditions of the input signal. This problem has been widely acknowledged in the community and is common not only for speaker recognition but for speech recognition, spoofing detection and other fields. The present state-of-the-art deep speaker embedding models are domain sensitive: they perform well in acoustic conditions that match that of the training data (in-domain), but degrade in mismatched acoustic conditions (out-of-domain). To deal with this problem ones use training data augmentation to improve the robustness of their systems in unknown domains. Others use in-domain data for adapting speaker embeddings to the specificdomain. These approaches include back-end adaptation [1, 2, 3] or more recent methods with adversarial training to learn condition-invariant deep embeddings [4, 5, 6, 7, 8].

This paper concentrates on the automatic tool for speech signal quality estimation. In the beginning, it was mainly considered as a tool for exploring the areas of applicability of different speaker verification methods and its adaptation techniques. However, it is also useful for preparing appropriate training data according to estimated acoustic parameters, for adaptation purposes during training or calibration.

One of the basic metrics for quality estimation is mean opinion score (MOS). It is widely used for service quality estimation in telecommunications and is a subjective quality evaluation measure, which means that it requires human assessors and is time-consuming.

Alternative objective metrics are based on a comparison of the original (reference) and encoded (distorted) signals, like Perceptual Evaluation of Speech Quality (PESQ) [9] and its improved version - Perceptual Objective Listening Quality Analysis (POLQA) [10]. Recent neural network approaches were trained to predict these estimates automatically [11]. In most applications the reference signal is unavailable and blind quality (without access to the reference or noise signal) estimation methods are in demand.

Over the past years, several methods for signal quality estimation that do not use reference samples have been proposed. These are neural network based approaches trained to predict the MOS, PESQ and similar measures on some sets of training data [12, 13]. Nevertheless, MOS, PESQ and POLQA measures are difficult to interpret and do not characterize which signal parameters affect the final measurement value. In the signal processing field, the quantitative characteristics of SNR and RT60 are commonly used to evaluate a particular distortion.

A comprehensive comparison of algorithms for blind RT60 and direct-to-reverberant energy ratio (DRR) estimation was facilitated by the Acoustic Characterisation of Environments (ACE) challenge [14]. The baseline algorithm and the best performed algorithm from [15] use sub-band analysis and maximum likelihood estimation for RT60.

The main idea of the current research was to develop a single method for automatic quality estimation, which allows to evaluate short-time (2 sec of speech) characteristics such as SNR and RT60. Section 3 presents a new neural network based approach for blind speech signal quality estimation in terms of SNR and RT60. Additionally, it is able to classify the type of underlying additive noise.

Since the described approach is performed for speech signal quality estimations, it uses only segments selected by voice activity detector (VAD) at the preliminary step. This research reveals the need for an accurate VAD that performs well in both clean and noisy environments and is able to generalize under unseen environments. Thus a novel neural network based VAD (Section 2) was invented for these purposes.

The proposed algorithms allow to perform the analysis of commonly used datasets like NIST, SITW, VOiCES, etc. from the new point of view and consider some calibration steps to improve speaker verification quality on them (Section 4).

2. Voice Activity Detection

Voice activity detection is one of the building blocks of speech processing systems. It selects features corresponding to speech segments before passing them to speaker verification, speech

http://dx.doi.org/10.21437/Interspeech.2020-1826

recognition or other methods. In this investigation the role of an accurate VAD procedure was reconsidered. Speech activity detection in case of varying acoustic conditions is a challenging task for classical energy-based VAD, especially in the presence of specific noises and distortions [16]. Therefore this paper presents the neural network based VAD trained to perform reliable results in case of noise and reverberation. Deep learning approaches have better modeling capabilities than traditional methods and have already achieved superior results in the VAD task [17, 18, 19, 20, 21, 22, 23].

2.1. System description

This work adapts the U-net [24] architecture to the speech activity detection task. Such architecture was originally introduced in biomedical imaging for semantic segmentation in order to improve precision and localization of microscopic images. U-net is a convolutional network [25] based on the deconvolutional idea [26]. In a deconvolutional network, a stack of convolutional layers, where each layer halves the size of the image but doubles the number of channels, encodes the image into a small and deep representation. That encoding is then decoded to the original size of the image by a stack of upsampling layers.

The proposed U-net based VAD uses a reduced version of the original architecture. It was firstly presented in the context of speaker verification task of the VOiCES Challenge [27]. Input features are 8kHz 23-dimensional Mel Frequency Cepstral Coefficients (MFCC) extracted for 25ms frame every 20ms. Half overlapping 2.56sec sliding window and 1.28sec overlap are used.This results in 128×23 input features size for the neural network. The aim of the neural network is to predict the 128 dimensional speech activity mask for every 2.56sec segment. Thus the resolution of the proposed speech detector is equal to 20ms. The final decoder layer is a global average pooling layer with sigmoid activation. Its output is used as the speech activity mask.

To train the network, a combination of binary cross-entropy loss function and dice loss [28] is used. The latter aims to maximize the dice coefficient between predicted binary segmentation set and ground truth binary labels.

2.2. Experiments

The U-net model was trained on NIST2008 and Russian speech subcorpus RusTelecom [29]. Augmentation process included reverberation and additional noise of different types. It was performed in the similar manner to augmentation for quality estimation described further in section 3.2. Speech labels were obtained from the oracle manual segmentation or Automatic Speech Recognition (ASR) [30] based VAD processed for clean version of the data.

To evaluate the performance of the proposed method we used NIST metrics FA, E_{miss}, $Precision$ and $Recall$ as implemented in `pyannote.metrics` [31]. Evaluation was done for hard conditions presented in the evaluation subset [32] from the DIHARD Diarization Challenge, since it represents the real life scenarios and has oracle VAD segmentation. Results from Table 1 confirm the improvement in terms of all used metrics. Furthermore, the improvement obtained by speaker verification system based on this VAD model confirms its robustness in challenging conditions presented in the VOiCES dataset [27].

Table 1: *Evaluation on DIHARD eval dataset.*

VAD	FA	E_{miss}	Precision	Recall
Energy based	4.59	41.61	95.12	58.39
U-net	4.19	23.85	98.14	76.15

3. Automatic Quality Estimation

Aimed to develop single system for multiple quality parameters estimation, we focused on the perspective deep learning approach, where data preparation plays a crucial role. Many papers underline the scarcity of data available for training deep models. Therefore, in this work the great attention was paid to the rigorous data augmentation procedure.

3.1. System description

The proposed models are trained in a multitask mode: one neural network is simultaneously trained to predict SNR, RT60 and background noise class. This is achieved through the use of three heads in the architecture of the neural network and three cost functions.To automatically evaluate SNR and RT60, the model is trained as a regressor and the mean squared error (MSE) loss is used. Automatic estimation of noise class is based on the use of a classifier and is trained using binary cross-entropy (BCE). During the training, a combined weighted loss function $\mathcal{L}$, with weights used for scaling, is used:

$$\mathcal{L} = 0.001\, MSE_{RT60} + MSE_{SNR} + 10\, BCE_{noise} \quad (1)$$

Three different architectures were compared in terms of estimation quality, speed and number of parameters:

- U-net - similar to the VAD architecture described above;
- FatCNN - convolutional neural network containing 5 convolutional layers. In the first layer 128 filters was used, followed by 2 × 2 max pooling and batch normalization layer. Other layers have, respectively, 256, 128, 128 and 512 filters, followed by 2 × 2 max pooling. We used ReLU as an activation function within the hidden units. Six fully-connected layers are then used prior to the output unit;
- ResNet18 - modification of well-known residual neural network with 8 ResNet blocks [33].

All architectures have the same high-level structure: deep quality embeddings of size 512 obtained after the global average pooing layer are used as an input to three linear layers - one per each head. For the "noise" head additional two-layer classification with softmax activation for 79 classes is used. Deep quality embeddings are further called q-vector (quality vector) as a sign of respect to deep speaker representation analogues.

We also investigated the influence of feature type and its resolution analysing 23 dimensional MFCC and 64 dimensional log mel filter bank (Fbank) features. All features are used without feature normalization.

3.2. Data preparation

Training of the quality estimator needs an dataset balanced according to the values of SNR, RT60 and types of noise. A special toolkit that allows to augment signal with specific acoustic parameters was developed for this purpose. Different bases in telephone and microphone channels, including NIST2002, NIST2008 and private STC databases, were taken as the sources

of "clean" speech signals. Stationary noises of 79 different types (factory, rain, babble, keyboard etc.) from Freesound [34], MUSAN [35] and manually collected from the Internet resources were used for noise augmentation. To obtain correct SNR value, in contrast to Kaldi implementation, power of signal was calculated only on speech segments after applying VAD, described in section 2. Reverberation effect was produced using the room impulse response (RIR) generator based on [36]. Four different RIRs were generated for each of 40,000 rooms with varying positions of speech and noise sources. It should be noted that, in contrast to the standard augmentation scenarios (Kaldi), both speech and noise signals were reverberated. In this case different RIRs generated for one room were used for speech and noise signals respectively. Thus more realistic data augmentation was obtained. We have already used this approach in our previous studies [37, 27].

In order to achieve a variety of acoustic conditions, we divided each sample into 1-minute segments and augmented each of them individually using randomly selected room and noise.

As a result of augmentation, the SNR range [-20; 30] dB and RT60 range [0.0045; 1.87] sec were evenly covered, which is critically important for the training process.

During the set of experiments the key factor in the success of effective SNR estimation training was revealed: it is highly important to use short-term SNR value for each 2-sec segment, but not the long-term value for the whole signal. Human speech and noises in real life are non-stationary signals. That is the reason why the global SNR value used for augmentation can not be interpreted as training label and needs to be re-estimated for each segment independently. To estimate local short-term SNR values for each training sample we found coefficients that are used in a linear combination of noise and source signal after reverberation, that for discrete signal can be expressed as

$$X_{aug}(i) = \alpha X^{rev}_{src}(i) + \beta X^{rev}_{noise}(i) \quad \text{for } i \in \{1, \dots, n\}$$

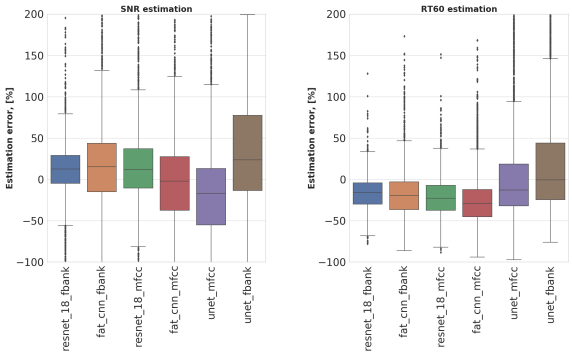
This can be done by solving the system of non-linear equations. After, the local SNR can be found by the following formula using these coefficients, and signal and noise energies $E^{rev}_{src}, E^{rev}_{noise}$:

$$SNR_{local} = 10 \log \left(\frac{\alpha^2 E^{rev}_{signal}}{\beta^2 E^{rev}_{noise}} \right)$$

3.3. Experimental results

For evaluation of the proposed methods we used the validation set contained the augmented version of the NIST2008 dataset (1000 files with 59-sec mean duration) that was not used during training and single-channel Eval part of the ACE dataset from [38] (4500 files with 19.4-sec mean duration).

Table 2 presents the comparison of all the proposed models with different features in terms of model sizes, speed for CPU (Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz) and quality on validation and test datasets. Predictably, the model quality increases with the number of parameters it uses, thus the better results were obtained for ResNet18 models, which are the heaviest among the proposed models and the worst results were obtained for the lightest models U-net. Notably, according to the minimum loss values on the validation set, systems based on Fbank features outperformed MFCC-based ones. However, results obtained for the ACE dataset can not confirm this statement: the best quality in terms of absolute errors of SNR predictions was achieved for the ResNet MFCC system, but Fbank based system with the same architecture showed the minimum



(a) (b)

Figure 1: *Representation of the estimations for all proposed models*

absolute errors of RT60 predictions. The accuracy of noise type recognition on the validation dataset for our best model reached only 26.26%. This can be explained by the large number of classes that are not discriminative enough: and overlap of properties between some classes can appear (for example, white Gaussian noise and rain sounds).

The representations of the models errors can be seen in boxplots in Figure 1. It displays the median and dispersion of the SNR and RT60 estimations errors for the ACE Eval dataset.

The best ResNet18 Fbank system was also evaluated in terms of ACE Challenge metrics: mean squared error (MSE), bias, and Pearson correlation coefficient of estimator implementations compared to the ACE corpus labels. The proposed system demonstrated 0.034 MSE, -0.118 Bias and 0.913 correlation coefficient, which is comparable to the published results of the ACE winner [15], Microsoft system [39] and CNN solution from [40].

We analyzed acoustic conditions of commonly used datasets in speaker verification and speech recognition tasks in different channels from the well-known challenges of the last years. In the microphone case we used SITW (development and evaluation parts), VOiCES (development and evaluation parts), JANUS Multimedia Dataset from NIST2019 and CHiME 5 datasets, and in the telephone case – NIST2019 development and evaluation parts. The distributions of SNR and RT60 values for our best ResNet18_Fbanks system (according to the loss value) are presented in Figure 2 for these datasets. Obtained results meet the subjective expectations and dataset descriptions. It is interesting to note that quality parameters distributions are almost the same for both development and evaluation sets of the NIST SRE 2019 and SITW benchmark but differ for the VOiCES challenge data parts. This accords with the previous studies that show SR system performance degradation for the VOiCES eval protocol compared to results obtained for the development protocol [27].

4. QE application for speaker recognition task

The major adverse conditions causing speaker recognition systems degradation are different channel and noise environments. The mismatches can exist in all combinations between training, testing and enrollment conditions. These mismatches lead to scores distributions scaling and shifting for different condi-

Table 2: *Comparison of the proposed systems in terms of quality estimates for ACE eval dataset, size of the model and speed on CPU*

Model	Features	Parameters	Size(Mb)	Speed (sec/sample)	Min loss on validation	Acc_{noise}, [%]	Part of ACE eval [%] with $\|Error_{SNR}\|<$			Part of ACE eval [%] with $\|Error_{RT60}\|<$		
							3db	5db	10db	100ms	200ms	300ms
U-net	MFCC	**1,398,724**	**5.34**	**0.0063**	144.35	14	32	46	70	36	55	**87**
	Fbank	1,530,180	5.84	0.0083	98.69	21	26	46	63	32	54	82
FatCNN	MFCC	3,516,168	13.41	0.0122	117.75	18	43	63	90	27	59	75
	Fbank			0.0222	83.71	25	46	67	92	44	68	82
ResNet18	MFCC	10,090,467	38.49	0.0593	99.00	21	**48**	**79**	**95**	42	69	81
	Fbank			0.1294	**68.55**	26	41	65	93	**51**	**78**	**87**

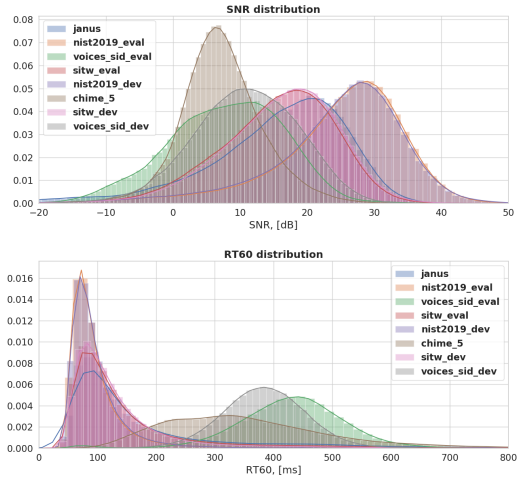


Figure 2: *Distributions of SNR (upper) and RT60 (lower) values obtained by ResNet18_Fbank for public datasets*

tions. Such scores instability impairs SR systems in real applications. In order to deal with this problem different calibration strategies [41, 42] and compensation techniques are implemented [43, 44].

Quality Measure Functions (QMF) [43, 45] are often used for score stabilization and calibration according to various enrollment and test duration conditions. Following [45], we used a similar approach to various speech quality conditions. For this purpose we applied our best blind speech quality estimation model (Resnet18 on Fbank features) for automatic SNR and RT60 assessment of enrollment ($snr_e, rt60_e$) and test ($snr_t, rt60_t$) speech segments. We proposed to use simple model to compensate speech condition scores shifting:

$$\begin{aligned} S_{new}(e,t) &= S_{raw}(e,t) - \delta_{qmf}(q_e,q_t) \\ \delta_{qmf}(q_e,q_t) &= C_t\mu_{tar}(q_e,q_t) + C_i\mu_{imp}(q_e,q_t) \end{aligned} \quad (2)$$

where $S_{raw}(e,t)$ and $S_{new}(e,t)$ are raw and new shift compensated verification scores, respectively, $\mu_{tar}(q_e,q_t)$ and $\mu_{imp}(q_e,q_t)$ are functions of the means of target and impostors scores distributions, respectively, which depends on enrollment speech segment quality (q_e: snr_e or $rt60_e$) and test speech segment quality (q_t: snr_t or $rt60_t$) estimations, C_t and C_i are tunable coefficients. To train the proposed quality shift compensation model we generated development protocols to estimate the means of target and impostors scores for particular q_e and q_t condition regions. The development protocols for different q_e and q_t regions were generated using the CHiME-5 dataset.

In our experiments we fitted second order polynomial models on the development estimations of μ_{tar} and μ_{imp} to obtain $\mu_{tar}(q_e,q_t)$ and $\mu_{imp}(q_e,q_t)$ functions approximations. Finally, the parameters C_t and C_i were tuned on pulled condition CHiME-5 verification protocol according to the best SR system performance. For our SR experiments we used the Resnet34-based extractor (ResNet34-MFB80-AM-TrainData-II) described in detail in [27].

Table 3 shows the experimental results of applying QMF-based shift compensation to SITW and VOiCES speaker verification evaluation protocols both for development and evaluation datasets. It can be seen from these results that simple QMF shift compensation allows to improve the SR system performance for all protocols. It should be noted that the most fruitful improvement was achieved for the challenging VOiCES eval protocol (see Figure 2).

Table 3: *Speaker recognition evaluation results on SITW and VOiCES protocols in terms of EER[%]/minDCF($P_{tar} = 0.01$)*

QMF type	VOiCES		SITW	
	dev	eval	dev	eval
Raw scores	1.24/0.197	6.02/0.430	2.00/0.193	2.17/0.216
SNR	1.20/0.190	5.44/0.415	1.86/0.183	**2.07/0.210**
RT60	1.25/0.197	5.54/0.401	2.01/0.193	2.16/0.216
SNR&RT60	**1.20/0.189**	**5.19/0.398**	**1.86/0.182**	2.07/0.214

5. Conclusion

We proposed new deep convolutional neural network based systems and their training schemes to build accurate voice activity detection model and blind short-term speech segments quality estimators. These models perform well in both clean and noisy unseen environments. The results obtained on the ACE challenge dataset show that multitask mode can be effective in training the predictor for both SNR and RT60 values using single deep CNN-based model. Additionally, we analyzed different commonly used speaker recognition benchmark datasets in terms of SNR and RT60 quality estimations. We managed to show that applying the QMF model with blind SNR and RT60 quality estimations of test and enrollment speech segments helps to improve speaker recognition system performance by using simple verification scores shift compensations.

6. Acknowledgements

This work was partially financially supported by the Government of the Russian Federation (Grant 08-08) and by the Foundation NTI (contract 20/18gr) ID 0000000007418QR20002.

7. References

[1] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for i-vector based speaker recognition," 2014.

[2] N. Brümmer, A. McCree, S. Shum, D. Garcia-Romero, and C. Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Odyssey*, 2014.

[3] M. J. Alam, G. Bhattacharya, and P. Kenny, "Speaker verification in mismatched conditions with frustratingly easy domain adaptation," 2018.

[4] J. Rohdin, T. Stafylakis, A. Silnova, H. Zeinali, L. Burget, and O. Plchot, "Speaker verification using end-to-end adversarial language adaptation," 2018.

[5] Q. Wang, W. Rao, S. Sun, L. Xie, E. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," 2018.

[6] X. Wang, L. Li, and D. Wang, "Vae-based domain adaptation for speaker verification," 2019.

[7] P. S. Nidadavolu, S. Kataria, J. Villalba, and N. Dehak, "Low-resource domain adaptation for speaker recognition using cycle-gans," 2019.

[8] Z. Meng, Y. Zhao, J. Li, and Y. Gong, "Adversarial speaker verification," *ICASSP*, 2019.

[9] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq): A new method for speech quality assessment of telephone networks and codecs," vol. 2, 2001.

[10] J. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i-temporal alignment," *Journal of the Audio Engineering Society*, vol. 61, pp. 366–384, 2013.

[11] G. Mittag and S. Möller, "Full-reference speech quality estimation with attentional siamese neural networks," in *ICASSP*, 2020.

[12] R. Dubey, "Non-intrusive objective speech quality assessment using features at single and multiple time scales," Ph.D. dissertation, 2014.

[13] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," 2018.

[14] J. Eaton, N. Gaubitch, A. Moore, and P. Naylor, "Estimation of room acoustic parameters: The ace challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, 2016.

[15] T. de M. Prego, A. A. de Lima, R. Zambrano-López, and S. L. Netto, "Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition," in *WASPAA*, 2015, pp. 1–5.

[16] M. Sahidullah and G. Saha, "Comparison of speech activity detection techniques for speaker recognition," *ArXiv*, vol. abs/1210.0297, 2012.

[17] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with lstm recurrent neural networks and application to holliwood movies," 2013.

[18] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on youtube using deep neural networks," in *INTERSPEECH*, 2013.

[19] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," *ICASSP*, pp. 2519–2523, 2014.

[20] G. Gelly and J.-L. Gauvain, "Optimization of rnn-based speech activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

[21] D. Augusto, J. Stuchi, R. Violato, and L. Cuozzo, *Exploring Convolutional Neural Networks for Voice Activity Detection*, 2017, pp. 37–47.

[22] Y. Jung, Y. Kim, Y. Choi, and H. Kim, "Joint learning using denoising variational autoencoders for voice activity detection," in *INTERSPEECH*, 2018.

[23] M. Lavechin, M.-P. Gill, R. Bousbib, H. Bredin, and L. P. Garcia-Perera, "End-to-end domain-adversarial voice activity detection," 2019.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI 2015*, Munich, Germany, 2015, pp. 234–241.

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[26] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *ICCV 2015*, 2015, pp. 1520–1528.

[27] A. Gusev, V. Volokhov, T. Andzhukaev, S. Novoselov, G. Lavrentyeva, M. Volkova, A. Gazizullina, A. Shulipa, A. Gorlanov, A. Avdeeva, A. Ivanov, A. Kozlov, T. Pekhovsky, and Y. Matveev, "Deep speaker embeddings for far-field speaker recognition on short utterances," 2020.

[28] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3DV 2016*, 2016, pp. 565–571.

[29] S. Novoselov, A. Shulipa, I. Kremnev, A. Kozlov, and V. Shchemelinin, "On deep speaker embeddings for text-independent speaker recognition," 2018.

[30] I. M. et al., "The STC ASR system for the VOiCES from a distance challenge 2019," in *INTERSPEECH*, 2019, pp. 2453–2457.

[31] H. Bredin, R. Yin, J. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," 2019.

[32] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The second dihard diarization challenge: Dataset, task, and baselines," 2019.

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

[34] G. R. Frederic Font and X. Serra, "Freesound technical demo," in *ACM International Conference on Multimedia*, 2013, pp. 411–412.

[35] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *ArXiv*, vol. abs/1510.08484, 2015.

[36] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[37] S. Novoselov, A. Gusev, A. Ivanov, T. Pekhovsky, A. Shulipa, G. Lavrentyeva, V. Volokhov, and A. Kozlov, "STC speaker recognition systems for the VOiCES from a distance challenge," in *INTERSPEECH*, 2019.

[38] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE challenge - corpus description and performance evaluation," in *WASPAA*, 2015.

[39] H. Gamper and I. Tashev, "Blind reverberation time estimation using a convolutional neural network," 2018.

[40] N. J. Bryan, "Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation," 2019.

[41] L. Ferrer and M. McLaren, "A discriminative condition-aware backend for speaker verification," 2019.

[42] A. Swart and N. Brummer, "A generative model for score normalization in speaker recognition," *arXiv preprint arXiv:1709.09868*, 2017.

[43] M. I. Mandasari, R. Saeidi, M. McLaren, and D. A. van Leeuwen, "Quality measure functions for calibration of speaker recognition systems in various duration conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2425–2438, 2013.

[44] A. Shulipa, S. Novoselov, and Y. Matveev, "Scores calibration in speaker recognition systems," in *SPECOM*, 2016.

[45] M. I. Mandasari, R. Saeidi, and D. A. van Leeuwen, "Quality measures based calibration with duration and noise dependency for speaker recognition," *Speech Communication*, vol. 72, pp. 126–137, 2015.