Non-Intrusive Estimation of Speech Signal Parameters using a Frame-based Machine Learning Approach

Dushyant Sharma*, Lucia Berger*, Carl Quillen* and Patrick A. Naylor†
† Department of Electrical and Electronic Engineering, Imperial College London, UK
* Nuance Communications Inc.
Email: dushyant.sharma@nuance.com

***Abstract*—We present a novel, non-intrusive method that jointly estimates acoustic signal properties associated with the perceptual speech quality, level of reverberation and noise in a speech signal. We explore various machine learning frameworks, consisting of popular feature extraction front-ends and two types of regression models and show the trade-off in performance that must be considered with each combination. We show that a short-time framework consisting of an 80-dimension log-Mel filter bank feature front-end employing spectral augmentation, followed by a 3 layer LSTM recurrent neural network model achieves a mean absolute error of 3.3 dB for C50, 2.3 dB for segmental SNR and 0.3 for PESQ estimation on the Libri Augmented (LA) database. The internal VAD for this system achieves an F1 score of 0.93 on this data. The proposed system also achieves a 2.4 dB mean absolute error for C50 estimation on the ACE test set. Furthermore, we show how each type of acoustic parameter correlates with ASR performance in terms of ground truth labels and additionally show that the estimated C50, SNR and PESQ from our proposed method have a high correlation (greater than 0.92) with WER on the LA test set.**

***Index Terms*—deep neural networks, clarity index, speech quality.**

I. Introduction

Recent years have seen great interest in the development of algorithms for the assessment of speech signal characteristics, including the level of reverberation [1], the classification of environmental sounds [2], bit rate [3] as well as estimation of the perceptual speech quality [4]. The ability to estimate such characteristics from a speech signal without a clean reference (non-intrusively) has applications in many speech processing tasks. These include audio forensics [5], hearing aids [6] and automatic speech recognition (ASR) [7]. Another application of such algorithms is for analysing speech data to understand the distributions of various acoustic parameters and use those distributions for the purpose of automating data augmentation and selection strategies for training robust ASR systems. Also recently it was shown how non-intrusive estimation of spatial parameters could be used for speaker change detection and diarization [8].

Reverberation has a significant impact on a speech signal by smearing temporal and spectral cues, flattening formant transitions, reducing amplitude modulations and increasing low-frequency energy [9]. The effects of reverberation can be modelled as convolution with a Room Impulse Response (RIR) [10] and although an RIR captures the overall characteristics of room reverberation, its estimation is complex for many practical applications. Instead a number of parameters encapsulating the RIR are used to characterize reverberation, the most common ones being the Reverberation Time (T60), Clarity Index (C50) and Direct-To-Reverberant Ratio (DRR). On the topic of non-intrusive estimation of reverberation parameters, a number of algorithms have been proposed and a recent IEEE challenge, the Acoustic Characterization of Environments (ACE) challenge, allowed evaluation of parameters on a standardized evaluation set [1]. It has been shown in [11] that room reflections arriving after roughly 50 ms of the direct path are perceived either as separate echoes or as reverberation. Parada. et. al. [12] further showed that, of the multitude of possible parameters, C50 was most correlated with ASR performance.

A data driven method for non-intrusive T60 and DRR estimation was presented in [13] and was the competition winner for the single channel DRR estimation task in the ACE challenge, achieving an Root Mean Square Error (RMSE) of 3.8 dB. The ACE challenge winner for T60 estimation was a method exploiting sub-band detection of free decay regions [14] with an RMSE of 0.254 s for blind T60 estimation. More recently, a Gammatone filterbank feature based approach using a Convolutional Neural Network (CNN) to estimate the T60 was shown to achieve an RMSE of 0.191 s on the ACE test set [15].

Estimating the perceptual quality of a speech signal in a non-intrusive manner is a challenging task due to the highly variable range of degradations encountered and the highly subjective nature of the task. Over the last decade a number of algorithms have been proposed for estimating speech quality, with some methods targeting the estimation of the mean opinion score (MOS) [16] directly, while others target the estimation of an intrusive metric such as Perceptual Evaluation of Speech Quality (PESQ) [17] and more recently, Perceptual Objective Listening Quality Analysis (POLQA) [18]. A recent approach [19] explores the use of Constant Q Transfom (CQT) and Mel Frequency Cepstral Coefficients (MFCC) features with a 2D CNN and a three layer Feed Forward Deep Neural Network (FFDNN) to estimate the MOS. They

report an RMSE of 0.42 MOS on their test set. In [20], [21] a classification and regressions tree (CART) model was trained using a number of long and short-term speech features to predict the PESQ score non-intrusively with an RMSE of 0.49 [20]. An extension of that method uses a mix of modulation domain features and MFCC with a recurrent neural network (RNN) model to predict jointly the voice activity detection (VAD) posterior and POLQA in short term segments of length 300 ms. This method was shown to have an RMSE of 0.29 POLQA with good generalization across languages not seen during training. A CNN based approach using a joint classification and regression network working on the magnitude log frequency domain was recently shown to have an RMSE of 0.28 with pesq [22] and works at an utterance level.

In this paper, we propose a multi-task machine learning framework for non-intrusive acoustic parameter estimation that includes voice activity detection, C50, PESQ and segmental SNR. While most recent methods for non-intrusive speech quality and reverberation parameter estimation operate on a large temporal window size (for example 4 seconds for T60 estimation in [15]) or at an utterance level [22], our framework is able to reliably estimate a number of parameters in short windows of length 300 ms using the multi-task training paradigm. The ability of estimating these parameters in short-time windows allows the use of these parameters in speaker diarization, signal quality assurance and for automatic data selection for ASR.

The remainder of the paper is organised as follows. In Section II we present the proposed framework followed by a description of the data sets and evaluation metrics in Section III. We finish with results in Section IV and conclusions in Section V.

II. Methods

This work builds on the non-Intrusive POLQA estimation model [4] which jointly predicts POLQA and VAD in short-time frames of 300 ms [4]. In this paper, we explore the joint prediction of C50, segmental SNR, PESQ and VAD using this framework. We also explore different feature extraction front-ends and machine learning based back-ends. In Section II-A we present the three different set of features and in Section II-B we present the two deep learning architectures explored in this paper.

A. Feature Extraction

1) MMF: The MMF feature set consists of the combination of MFCC and MDCC features as described in [4] along with an additional feature that models the variation in the fundamental frequency [23]. MFCCs are a common feature set for a number of speech processing algorithms such as ASR [24]. In our method, we use a sample rate of $f_s = 16\,\text{kHz}$ with a pre-emphasis factor of 0.97 and extract 31 MFCC features every 10 ms using a 25 ms analysis window. These are appended to 231-dimensional MDCC features, which were first presented in [4] and model the modulation information in the signal. In this work we use a larger number of MDCC coefficients (231) than in our previous work by selecting a larger upper triangle from the DCT stage. Finally, a 7 dimension feature vector modelling the fundamental frequency variation is appended to this set.

2) MFB: A second type of feature extraction we explore provides Mel Filterbank Coefficients. These are similar to the MFCC features described in the MMF section, using the same signal processing with the exception of the DCT and instead of 31, 80 Mel Filterbanks are used. In this work we explore an additional regularization strategy allied to the MFB features whereby time and frequency blocks are zeroed or dropped in a randomized manner [25]. For our application, given that we are working with short-time blocks, we only apply the frequency dropping spectral augmentation here, chosen to randomly drop 3% of the Mel Channels for a given block of frames.

3) PASE: In addition to the well established spectral features discussed before, we explore a recent feature set that is based on a machine learning approach of extracting directly from the waveform a Problem-Agnostic Speech Representation (PASE) of speech using a self-supervised encoder-discriminator [26]. In this paper we explored this feature set as provided by the original authors, without further retraining for our particular task.

B. Deep Learning Architectures

We experimented with two different deep learning architectures as described in the following section. In training both of these models, VAD is estimated jointly with the other acoustic parameters and the models actually predict the average VAD posterior across the frames in the context of interest, defined here as mean VAD Posterior (VADP). Here we define the context size to be the duration of the segment of a signal that is used for performing the regression modelling. During the inference phase, the VADP can be used as a confidence measure and used to prune the estimated scores. In the following, we only consider segments that pass a VADP threshold of 0.5 or greater.

In both the learning architectures explored here, we use an RMSE cost function during training with an Adam [27] optimizer and a decaying learning rate schedule. The RNN and CNN models have 125,880 and 16,802 trainable parameters, respectively.

1) LSTM: The first deep learning architecture we explore is a recurrent structure based on long short term memory (LSTM) cells [28], which has been shown to be a powerful architecture for modelling time varying features, such as those used in speech signal processing. The LSTM is composed of an input layer and three hidden layers, arranged in a $108 \times 54 \times 27$ cell topology (for each time step).

2) CNN: The second deep learning architecture we explore is a compact CNN model based on SwishNet [29], which was originally proposed for VAD modelling and itself inspired by the WaveNet [30] model. This network applies 1-dimensional filters across time to the input features and makes use of various enhancements such as dilated convolutions and gated

activations to improve the modelling power. The selected model has eight convolutional layers, which apply causal convolutions, with sigmoid and tanh activations. The original system was trained with MFCC features but here we explore the use of MFB features with this model and experimented with different filter sizes and added a dropout layer in the architecture. Additionally, the output layer is modified to support a regression task using a linear output layer with four nodes.

III. Data and Metrics

A. Training and Development Data

The training data is based on speech from the clean training partition of the Wall Street corpus [31]. This forms the base material for the training data set and is artificially corrupted by convolution with RIRs generated using the Image method [32] and additive noise (ambient, babble, domestic and white). The RIRs are sampled with a T60 in the range [0.1 to 1.0 s], C50 in the range [0 to 30 dB] and DRR in the range [-15 to 5.0 dB]. An exponential decay is applied to some of the RIRs to create a larger set of RIRs with a high C50. The noise data used in training is part of internal as well as open source collections and are different from those used in the test sets.

B. Test Data

We use two test datasets for evaluating the performance of the proposed algorithms as described below. In both cases the base speech material is from a different source to the training data (and thus has no overlap in language, speaker or recording system) and the RIRs and noise sources are also completely separate with no overlap.

The first is the ACE Challenge test-set [1]. In this work, we consider the single channel test set and estimate the full-band C50 metric. Please note that the ACE Challenge test set did not include C50 estimation criteria and this was added as an extra label following the definition in [12].

The second is the Libri Augmented (LA) test set, which is based on speech from the Libri clean dataset and reverberation and noise are added artificially. The reverberation conditions are a sample of all available RIRs in the ACE data as well as measured RIRs from the Aachen database [33]. The noise types included in this set are ambient, babble, household and white, added to the speech in an SNR range of -5 to 30 dB. This set is asymmetrically corrupted to allow a wide range of degradations to be evaluated and the use of Libri speech data allows us to decode with a Libri speech trained ASR system.

C. Evaluation Metrics

In the following, P_e and P_t are the estimated and true scores (for the acoustic parameters – C50, SNR and PESQ) and the error in estimating a sample is defined as $E(n) = P(n)_e - P(n)_t$. For evaluating the VAD performance, the posterior generated during inference is mapped to a binary decision by thresholding at a heuristic value of 0.5, thus allowing the VAD performance to be treated as a classification task as its performance measured using the F1 score as defined in following table.

Metric	Description
Pearson Correlation Coefficient (R)	$\mathrm{R}= \frac{\sum_{n=1}^{N}(P_e(n)-\bar{P_e})(P_t(n)-\bar{P_t})}{\sqrt{\sum_{n=1}^{N}(P_e(n)-\bar{P_e})^2}\sqrt{\sum_{n=1}^{N}(P_e(n)-\bar{P_e})^2}}$
Root Mean Square Error (RMSE)	$\mathrm{RMSE} = \sqrt{\sum_{n=1}^{N}\frac{1}{n}E(n)^2}$
Mean Absolute Difference (MAD)	$\mathrm{MAD} = \sum_{n=1}^{N}\frac{1}{n}\lvert E(n)\rvert$
F1 Score	$\mathrm{F1} = 2.\frac{\mathrm{precision.recall}}{\mathrm{precision+recall}}$
Word Error Rate (WER)	$\mathrm{WER} = \frac{\mathrm{Substituions+Deletions+Insertions}}{\mathrm{Substituions+Deletions+Correct}} \times 100$

In this paper, we also evaluate the correlation of various metrics with the End-to-End ESPnet ASR system [34] trained on the 1000 hr Librispeech corpus (the full training partition - this partition is completely separate from the LA test set described in the previous sub-section in terms of speakers and text). The performance of an ASR system is typically evaluated in terms of Word Error Rate (WER) as defined in the following table.

IV. Results

In this section we evaluate the performance of the five systems, representing four types of feature front-ends and two types of regression model back-ends. Table I shows the results for C50 estimation on the ACE challenge test set. The MFB feature set with spectral augmentation and an LSTM model achieves the best performance, with an RMS error of 3.04 dB.

In table II we present the results for all the parameters that are jointly estimated by the different systems on the LA test set. It can be seen that the MFB features with spectral augmentation perform well on this test set, outperforming the others for PESQ estimation. Although the PASE feature set with the LSTM model achieves the lowest errors for C50 and SNR estimation on the LA test set, it comes with a high computational complexity (as it has almost 6 million trainable parameters - also increasing the risk of over-fitting to the training data) [1]. The mean absolute error for SNR estimation with the MFB is lower than 3 dB for all systems tested. The SNR results presented here are computed as the mean segmental SNR for each 300 ms segment in the entire test set. All systems tested on the LA test set performed well for joint VAD estimation, achieving an F1 score between 0.92 and 0.93. We also show in table III how the MFB with spectral augmentation and LSTM model performs for the four different

[1]Note that in this work, we do not re-train the PASE feature extraction (which might lead to more gains).

Feature Type	Spec. Aug.	Model Type	C50 (dB) RMSE	C50 (dB) MAE
MMF	✗	LSTM	3.44	2.74
MFB	✗	LSTM	3.15	2.57
MFB	✓	**LSTM**	**3.04**	**2.40**
PASE	✗	LSTM	4.25	3.40
MFB	✓	CNN	3.55	2.88

TABLE I
C50 ESTIMATION ON THE ACE TEST SET FOR DIFFERENT FEATURE FRONT-ENDS AND REGRESSION MODELS.

types of additive noise in the LA test set. It can be seen that the worst type of noise for joint estimation of C50, SNR and PESQ is babble noise, which is expected as babble is a speech like characteristic and moreover, in a short-time window of 300 ms this effect is likely to be more prominent.

As one of the motivations for this type of non-intrusive measure is in applications related to ASR, we present in table IV the correlations between WER and its various components with ground truth C50, T60, DRR, SNR and PESQ. The results presented here between the various acoustic parameters and WER are after 'binning' the parameters (4 dB for C50 and SNR and 0.7 for PESQ). We confirm that C50 is the highest correlated reverberation measure in addition to SNR and PESQ. In the bottom part of table IV we can see how the three estimated parameters from the MFB with spectral augmentation and LSTM model correlate with WER and other ASR metrics, all achieving a correlation higher than 0.92 with WER.

V. CONCLUSIONS

We presented a non-intrusive method that jointly estimates a number of interesting acoustic signal properties associated with a speech signal. These include estimation of the level of reverberation in the signal, the level of background noise, the perceptual signal quality an estimate of speech presence in short-time windows of 300 ms. The precise metrics that our method estimates thus includes C50, segmental SNR and PESQ. We show that a low complexity MFB based feature extraction with spectral augmentation with an LSTM model achieves good performance for all the parameters and provides a good trade-off in performance and complexity. We show the performance of this system on the ACE and LA test sets and show that it has a stable performance across different noise types.

Furthermore, we show how different ground truth and estimated parameters correlate with ASR performance (in terms of WER as well as the sub-components such as insertions, deletions etc.). We show, firstly that of the three reverberation level estimation parameters, C50 is the most correlated with WER, and in addition, SNR and PESQ are also highly correlated with WER. Lastly, we show that the C50, SNR and PESQ estimated by the NISA model are also highly correlated with ASR performance.

REFERENCES

[1] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 10, pp. 1681–1693, Oct. 2016.

[2] N. Davis and K. Suresh, "Environmental sound classification using deep convolutional neural networks and data augmentation," in *IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, 2018.

[3] U. J. D. Sharma and P. A. Naylor, "Non-intrusive bit-rate detection of coded speech," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2017.

[4] D. Sharma, A. O. Hogg, Y. Wang, A. Nour-Eldin, and P. A. Naylor, "Non-intrusive polqa estimation of speech quality using recurrent neural networks," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2019.

[5] S. Ikram and H. Malik, "Digital audio forensics using background noise," in *2010 IEEE International Conference on Multimedia and Expo*, 2010.

[6] F. Saki and N. Kehtarnavaz, "Automatic switching between noise classification and speech enhancement for hearing aid devices," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016.

[7] L. F. Gallardo, S. Moller, and J. Beerends, "Predicting Automatic Speech Recognition Performance Over Communication Channels from Instrumental Speech Quality and Intelligibility Scores," in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*. ISCA, Aug. 2017, pp. 2939–2943.

[8] M. Hu, D. Sharma, S. Doclo, M. Brookes, and P. Naylor, "Speaker change detection and speaker diarization using spatial information," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015.

[9] P. Assmann and Q. Summerfield, "The perception of speech under adverse conditions," in *Speech processing in the auditory system*, 2004.

[10] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," P. A. Naylor and N. D. Gaubitch, Eds. PUB-SV, 2010.

[11] A. Boothroyd, "Room acoustics and speech perception," in *Seminars in Hearing*, 2004.

[12] P. P. Parada, D. Sharma, J. Lainez, D. Barreda, T. van Waterschoot, and P. A. Naylor, "A single-channel non-intrusive c_{50} estimator correlated with speech recognition performance," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 24, no. 4, pp. 719–732, Apr. 2016.

[13] P. P. Parada, D. Sharma, T. van Waterschoot, and P. A. Naylor, "Evaluating the non-intrusive room acoustics algorithm with the ACE challenge," in *Proc. ACE Challenge Workshop, a satellite of IEEE-WASPAA*, New Paltz, NY, USA, Oct. 2015.

[14] T. d. M. Prego, A. A. de Lima, R. Zambrano-Lopez, and S. L. Netto, "Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2015.

[15] H. Gamper and I. J. Tashev, "Blind reverberation time estimation using a convolutional neural network," in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018.

[16] ITU_T_P10, "Amendment 2: Vocabulary for performance and quality of service amendment." INST_ITU_T, Recommendation, 2009.

[17] "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," INST_ITU_T, Recommendation P.862, Nov. 2003.

[18] "Perceptual objective listening quality assessment: An advanced objective perceptual method for end-to-end listening speech quality evaluation of fixed, mobile, and IP-based networks and speech codecs covering narrowband, wideband, and super-wideband signals," INST_ITU_T, Standard, Jan. 2011.

[19] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[20] D. Sharma, L. Meredith, J. Lainez, D. Barreda, and P. A. Naylor, "A non-intrusive PESQ measure," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec. 2014, pp. 975–978.

Feature Type	Spectral Augmentation	Model Type	C50 (dB) RMSE	C50 (dB) MAE	C50 (dB) R	SNR (dB) RMSE	SNR (dB) MAE	SNR (dB) R	PESQ RMSE	PESQ MAE	PESQ R	VAD F1	Trainable Parameters
MMF	✗	LSTM	3.69	2.88	0.71	2.30	1.78	0.87	0.41	0.34	0.82	0.92	125,824
MFB	✗	LSTM	3.96	3.11	0.75	2.79	2.20	0.83	0.39	0.33	0.87	0.93	125,824
MFB	✓	**LSTM**	**4.09**	**3.32**	**0.73**	**2.96**	**2.29**	**0.77**	**0.38**	**0.32**	**0.85**	**0.93**	**125,824**
PASE	✗	LSTM	3.62	2.82	0.76	1.78	1.36	0.92	0.41	0.36	0.87	0.93	5,943,844
MFB	✓	CNN	4.41	3.51	0.60	3.27	2.58	0.71	0.42	0.35	0.82	0.93	16,346

TABLE II
PERFORMANCE OF THE VARIOUS FEATURE FRONT-ENDS AND REGRESSION MODELS ON THE LA TEST SET. THE MFB FEATURE WITH SPECTRAL AUGMENTATION AND AN LSTM NETWORK (BOLD) PERFORMED BEST ON THE ACE TEST SET FOR C50 ESTIMATION.

Noise	C50 (dB)	SNR (dB)	PESQ
Ambient	4.15	3.25	0.36
Babble	4.40	3.39	0.42
Household	3.93	3.07	0.36
White	3.84	1.87	0.38

TABLE III
RMS ERROR ON THE LA TEST SET FOR C50, SNR AND PESQ ESTIMATION FOR EACH OF THE FOUR TYPES OF ADDITIVE NOISE CONDITIONS.

Parameter	WER	Cor.	Sub.	Del.	Ins.
GT-C50	-0.82	+0.64	-0.53	-0.63	-0.92
GT-T60	+0.75	-0.63	+0.63	+0.56	+0.97
GT-DRR	-0.78	+0.24	-0.49	-0.10	-0.50
GT-SNR	-0.94	+0.90	-0.93	-0.87	-0.87
GT-PESQ	-0.92	+0.87	-0.91	-0.61	-0.44
NISA-C50	-0.92	+0.98	-0.91	-0.86	-0.70
NISA-SNR	-0.98	+0.91	-0.95	-0.93	-0.21
NISA-PESQ	-0.92	+0.98	-0.17	-0.95	+0.26

TABLE IV
CORRELATION ANALYSIS OF VARIOUS ACOUSTIC PARAMETERS WITH WER USING THE LIBRI 1000 HR. TRAINED ESPNET MODEL. THE PREFIX GT REFERS TO GROUND TRUTH MEASURES AND THE NISA PREFIX REFERS TO THE ESTIMATED PARAMETERS BY THE MFB FEATURE WITH SPECTRAL AUGMENTATION AND AN LSTM MODEL.

[21] D. Sharma, Y. Wang, P. A. Naylor, and M. Brookes, "A data-driven non-intrusive measure of speech quality and intelligibility," *Speech Communication*, vol. 80, pp. 84–94, Jun. 2016.

[22] X. Dong and D. S. Williamson, "A classification-aided framework for non-intrusive speech quality assessment," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.

[23] K. Laskowski, M. Heldner, and J. Edlund, "The fundamental frequency variation spectrum," in *Proceedings of FONETIK*, vol. 2008. Citeseer, 2008, pp. 29–32.

[24] H. Hermansky, J. R. Cohen, and R. M. Stern, "Perceptual properties of current speech recognition technology," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 1968–1985, 2013.

[25] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*, 2019.

[26] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*, 2019.

[27] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[28] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[29] M. Hussain and M. A. Haque. (2018) Swishnet: A fast convolutional neural network for speech, music and noise classification and segmentation. [Online]. Available: https://arxiv.org/abs/1812.00149

[30] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[31] C.-I. (WSJ1). (1994) Complete ldc94s13a. dvd. philadelphia: Linguistic data consortium. [Online]. Available: https://catalog.ldc.upenn.edu/LDC94S13A

[32] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[33] S. LIEBICH, J. FABRY, P. JAX, and P. VARY, "Acoustic path database for anc in-ear headphone development," in *International Congress on Acoustics (ICA)*, 2019.

[34] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, Y. Enrique, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*, 2018.