INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

eSTImate: A Real-time Speech Transmission Index Estimator With Speech Enhancement Auxiliary Task Using Self-Attention Feature Pyramid Network

Bajian Xiang[1,†], *Hongkun Liu*[1,†,*], *Zedong Wu*[1], *Su Shen*[1], *Xiangdong Zhang*[1],

[1]Yóusonic Technology, Beijing, China

{bajian.xiang, hongkun.liu, zedong.wu, su.shen, xiangdong.zhang}@yousonic.tech

Abstract

The Speech Transmission Index (STI) is a crucial metric for evaluating speech intelligibility, but its standard measurement method is too complicated for real-time applications. Though recently proposed deep learning based STI estimation schemes can effectively address the problem, existing methods still fall short of covering all possible STI scenarios. This paper presents eSTImate: an end-to-end deep learning system for real-time STI blind estimation that integrates the tasks of STI estimation and speech enhancement through a feature pyramid auxiliary learning architecture and incorporates multi-head attention mechanisms. The proposed model demonstrates the performance of state-of-the-art, achieving a low mean absolute error of 0.016 and root mean square error of 0.021 on the constructed dataset that covers the whole range of STI, highlighting its potential to provide accurate and consistent real-time STI estimation across diverse real-world scenarios.

Index Terms: speech transmission index estimation, speech enhancement, deep neural networks, auxiliary learning

1. Introduction

As speech signal propagates, it is more or less modified by acoustic factors in the surrounding environment, such as ambient noise and reverberation characteristics, resulting in the degradation of speech quality. To assess the speech intelligibility at the listener's end, a range of evaluation methods have been suggested. Subjective evaluations, including MOS [1], involve human listeners' opinions about the speech quality. These methods remain as valuable tools in a wide range of contexts despite limitations like individual differences. Objective evaluations utilize algorithms and mathematical models to evaluate speech quality, including PESQ [2, 3], AI [4], STI [5, 6] and STOI [7]. Notably, the Speech Transmission Index (STI) has been demonstrated to exhibit strong correlation with speech intelligibility [8].

STI is quantified on a scale of 0 to 1, with higher values indicating better speech intelligibility. To provide a more specific and standardized evaluation, STI is divided into standard categories defined by five grades: bad, poor, fair, good, and excellent [9]. The differentiation criteria for these grades are set at 0.30, 0.45, 0.60, and 0.75, respectively.

The standard direct measurement method of STI employs 7-octave bands and 14 modulation frequencies to sequentially generate modulation test signals. This approach typically takes the handheld devices about 15 minutes for complete measurement and makes it difficult to be adopted in practice. As an alternative, the indirect method derives STI from the Room Impulse Response (RIR), offering a simpler way of computation [10]. However, the acquisition of the the RIR as a computational condition during real-time measurement is exceedingly challenging in practical scenario.

Due to recent advancements in neural networks, novel deep-learning-based algorithms have been developed for STI measurement with the potential to overcome the limitations of traditional measurement methods. Seetharaman et al. proposed a fully Convolutional Neural Network (CNN) to estimate the STI value directly from the input PCM audio [11]. While this network exhibited comparable performance to human in discriminating STI conditions, the range of STI values predicted by the model was constrained by the distribution of the underlying dataset, with a lower bound about 0.60. Additionally, Duangpummet et al. proposed a scheme that incorporates the Temporal Amplitude Envelope (TAE) into a CNN [12]. Despite the broadened range of measurements, a reduction in accuracy is observable when compared to the previous study. Afterwards, the authors employed an extended RIR model into a similar network structure to enable the simultaneous prediction of the STI and 5 other room acoustic parameters, yielding improved accuracy compared with their previous method [13]. To date, López et al. have presented a deep Convolutional Recurrent Neural Network (CRNN) for the blind estimation of the STI and 5 other room acoustic parameters, achieving a competitive result in STI in terms of both accuracy and estimation range [14].

However, the current methods are still unable to effectively distinguish between the lowest grades of STI, namely bad and poor scenario, as they fail to provide an accurate measurement for STI values below 0.30. Moreover, predicting STI simultaneously with other room parameters may not be a practical approach, as STI is not solely related to RIR but also affected by noise. Typically, estimating STI with other room parameters may impede real-time STI prediction, for STI is greater in temporal volatility than other room parameters.

This paper presents a novel auxiliary learning framework that integrates the tasks of blind STI estimation and Speech Enhancement (SE) through a Feature Pyramid Network (FPN) architecture [15]. To make better use of the correlations between speech sequences, a multi-head attention mechanism is incorporated into the network.To meet all possible needs of real measurement scenarios, a dataset covering the full range of STI is constructed. The proposed network exhibits SOTA performance, as evidenced by its achievement of a low Mean Absolute Error (MAE) of 0.016 and Root Mean Square Error (RMSE) of 0.021 on the dataset we built, showing the potential to estimate precise and consistent real-time STI estimation across diverse real-world scenarios.

†Equal Contribution
* Corresponding author

10.21437/Interspeech.2023-727

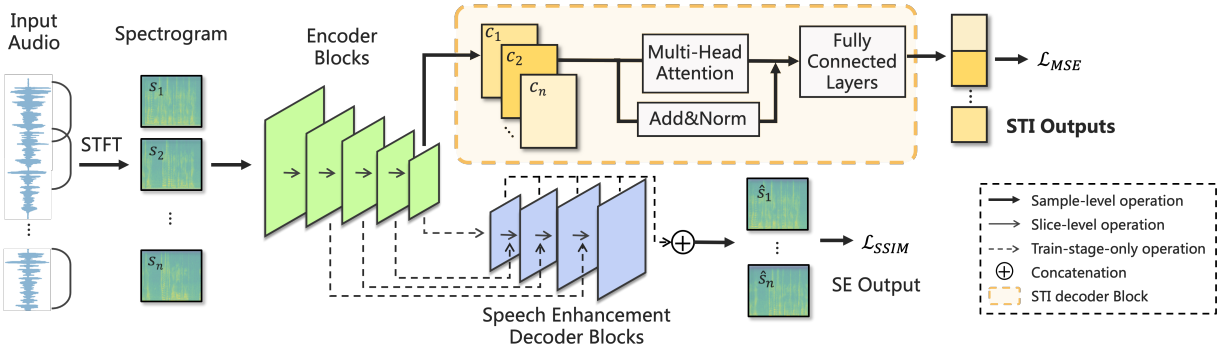

Figure 1: *Architecture of the proposed eSTImate system during the training phase.*

2. Problem formulation

In the time domain, an observed speech signal $y(t)$ captured by a microphone is a combination of the anechoic speech signal $x(t)$, the RIR $h(t)$, and the ambient noise $n(t)$, mathematically presented as

$$y(t) = x(t) * h(t) + n(t), \tag{1}$$

where $*$ denotes convolution operation. Typically, the standard indirect methods require all these 3 components to calculate STI. In this process, signal-to-noise ratio (SNR) is initially calculated for each octave band as

$$\rho_k = 10 \times log_{10} \frac{\int_0^\infty x_k^2(t)dt}{\int_0^\infty n_k^2(t)dt}, \tag{2}$$

where ρ_k, x_k and n_k are the SNR, original signal and ambient noise in octave band k, respectively. Afterwards, the Schroeder Method [16] is used to calculate the Modulation Transfer Function (MTF) with adjustment for noise in each octave band k and modulation frequency f_m, as

$$m_k(f_m) = \frac{|\int_0^\infty h_k^2(t)e^{-j2\pi f_m t}dt|}{\int_0^\infty h_k^2(t)dt}[1 + 10^{-\frac{\rho_k}{10}}]^{-1}, \tag{3}$$

where $h_k(t)$ is the RIR in octave band k. Then the effective signal-to-noise ratio N_{k,f_m} is calculated using MTF as

$$N_{k,f_m} = 10 \times log_{10} \frac{m_k(f_m)}{1 - m_k(f_m)}, \tag{4}$$

where N_{k,f_m} should be limited between -15 and $+15$. Next, by averaging over the 14 modulation frequencies, the Modulation Transfer Index (MTI) can be calculated as

$$M_k = \frac{1}{14} \sum_{m=1}^{14} \frac{N_{k,f_m} + 15}{30}, \tag{5}$$

where M_k represents the MTI value in octave band k. Afterwards, the STI value can be obtained using:

$$STI = \sum_{k=1}^{7} \alpha_k \times M_k - \sum_{k=1}^{6} \beta_k \times \sqrt{M_k \times M_{k+1}}, \tag{6}$$

where α_k is the gender-specific weight factor for each octave band, and β_k is the gender-specific redundancy factor between two adjacent octave bands.

For blind estimation of STI, the task is to derive the STI directly from the observed signal $y(t)$ without any supplementary information of $h(t)$, $n(t)$, and $x(t)$.

3. Proposed method

In this section, we present eSTImate, an end-to-end deep learning system for real-time STI blind estimation with speech enhancement as the auxiliary task. Figure 1 illustrates its main structure and workflow during the training phase. The network comprises three primary modules: an encoder using Resnet50 as the backbone, an STI decoder based on a multi-head attention mechanism [17], and a speech enhancement decoder based on feature pyramid hierarchy. The encoder, shared by 2 tasks, forms the STI estimation pipeline with the STI decoder and forms a Feature Pyramid Network (FPN) with the speech enhancement decoder for the auxiliary task of speech enhancement. The inclusion of the auxiliary task aims to help the shared encoder distinguish between noisy and clean speech, leading to improved feature representation and enhanced model generalization. Notably, the speech enhancement decoder can be omitted in STI estimation to enhance computational efficiency.

3.1. Raw audio encoding

To achieve real-time STI prediction, we introduce a speech slicing mechanism as a pre-processing step. Each input utterance is divided into frames of equal time intervals, with a certain length of overlap between every two adjacent frames, which ensures that the proposed system is adaptable to inputs of different durations. Moreover, the overlap scheme provides each frame with more contextual information and serves as the time interval for real-time prediction during the model inference phase.

For each utterance, the frames obtained by the above-mentioned slicing mechanism in the time domain are transformed into spectrogram features using Short-Time Fourier Transform with an FFT size of 512, resized to 224×224 using nearest neighbor interpolation, and normalized to the range of -1 and 1. Subsequently, the encoder blocks take the spectrogram features $S = (s_1, s_2, \cdots, s_n)$ as the input and continuously perform down-sampling, in which the size of adjacent features is halved, while the number of channels is doubled.

3.2. STI decoding

To better utilize the contextual relationship between the frames, a multi-head attention mechanism is introduced in the STI decoding phase. For each utterance, all the frames, after down-sampled by the encoder blocks, are concatenated at an additional dimension to form the deepest feature map $C = (c_1, c_2, \cdots, c_n)$, and then input to the STI Decoder. Each sample-level feature C is first processed by a multi-head self-

attention module, and the resulting features are added to C via residual connections and layer normalization. The features processed by the attention module are then mapped to the target domain of STI via a 5-layer fully connected network to obtain the STI prediction for each frame. The first 4 layers of the fully connected network contain linear layers with a halving feature dimension, ReLU activation, and batch normalization, with the features eventually reduced to 64. Dropout is also added to prevent overfitting. The last layer maps the feature directly to the STI prediction.

3.3. Speech enhancement decoding

The purpose of introducing speech enhancement as an auxiliary task is to help the network better understand the differences between clean and noisy speech during training. Specifically, during backpropagation, the speech enhancement auxiliary task adjusts the shared encoder to improve the overall generalization performance of the model. It should be noted that in the actual use of the eSTImate framework, we discard the speech enhancement decoder module and only use encoding to improve the system's prediction speed.

The speech enhancement decoder is designed to extract high-level semantic information from the encoder and perform bottom-up sampling. It first reduces the channel dimension of the lowest-level feature map obtained by down-sampling from the encoder. Then, it adds the up-sampled and 1×1 convolved lowest-level feature map to the corresponding high-level feature map to generate a new set of feature maps. Finally, all the feature maps generated during the up-sampling process are fused and up-sampled to the same size as the input spectrogram, forming the final output of estimated clean speech $\hat{S} = (\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_n)$.

3.4. Training objective

The Mean Square Error (MSE) between the predicted STI and the ground truth STI, as well as the Structural Similarity Index (SSIM) [18] between the predicted clean spectrogram and the ground truth clean spectrogram, are used as the training objectives, as

$$\mathcal{L} = \alpha \times \mathcal{L}_{MSE}(\hat{STI}, STI) + \beta \times \mathcal{L}_{SSIM}(\hat{S}, S) \quad (7)$$

where α and β are weights for the MSE and SSIM losses, respectively. The MSE loss measures the average squared difference between the predicted and true STI values, while the SSIM loss assesses the structural similarity between the predicted and true clean spectrograms, taking into account luminance, contrast, and structural information.

During the training process, an automatically weighted loss technique for multi-task learning framework is used to guide the selection of α and β [19].

4. Experiments

4.1. Dataset

The dataset built for training the proposed eSTImate system consists of $592,928$ synthesized single-channel noisy audio samples, with a sampling rate of $16\,\mathrm{kHz}$. Each utterance has a duration of $20\,\mathrm{s}$ to $30\,\mathrm{s}$ and is accompanied by its corresponding real-time STI ground truth value, as well as a clean audio aligned with the noisy audio. The proposed dataset covers the entire range of STI values from 0 to 1, and the distribution is visualized in Figure 2.

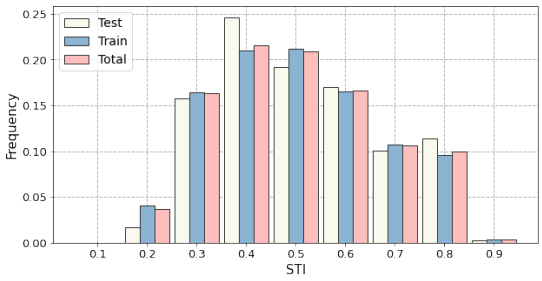

Figure 2: *Frequency distribution of the proposed STI dataset.*

A total of 108 real-world RIRs are used in our experiments. Among them, 32 RIRs were recorded in different offices in China using a standard recording method, while the remaining 76 RIRs were obtained from the EchoThief Impulse Response Library [20]. The selected RIRs cover a range of Reverberation Time (T60) from 0.4 to 1.0, with 18 RIRs at each 0.1 interval, measured using the Dirac tool from Acoustics Engineering. For each interval, we randomly selected 2 RIRs to generate test data.

The TIMIT dataset [21] was used as the clean speech corpus, which contains recordings of 630 individuals from 8 major dialect regions in the United States each speaking 10 standard sentences [22]. The noise used was primarily sourced from the TUT Acoustic Scenes dataset [23], which consists of audio recordings of various street environments, including 6 different categories such as cars, braking sounds, and pedestrians [24]. Additionally, we added pink noise and babble noise to our noise library. For each reverb speech, we added 3 types of noise randomly, with each type of noise added at 3 different SNR levels of 0dB, 6dB, and 20dB, respectively.

4.2. Experimental Details

In training, the batch size is 7. The initial learning rate is 10^{-5}, halved every 10 epochs. The inference time for each $4\,\mathrm{s}$ frame is $0.1\,\mathrm{s}$ and $0.047\,\mathrm{s}$ respectively with AMD Ryzen5 5600H CPU and 24G GeForce RTX 3090 GPU.

4.3. STI evaluation

The performance of the proposed eSTImate algorithm was evaluated on a test set with STI ground truth values ranging from 0.21 to 0.95. Data analysis was conducted at intervals of 0.01, and the results are shown in Figure 3. The red dashed line in the figure represents the ideal prediction, while the blue solid line represents the mean prediction of the eSTImate model on the test set, and the blue shaded area represents the standard deviation of the prediction. The results demonstrate a strong fitting performance with very small prediction errors, achieving an RMSE of 0.021 and an MAE of 0.016.

The model shows the closest prediction results to the ground truth values within the range of STI values from 0.3 to 0.8, with a slight overestimation in the range of 0.2 to 0.3 and underestimation above 0.8. This prediction pattern is strongly related to the distribution of our dataset, where there are relatively few samples with STI values below 0.3 and above 0.8, causing the model's predictions for these ranges to be less accurate. Although there is a slight deviation in the prediction results at both ends of the range, it has little impact on the STI classification task in practical applications. To demonstrate this, we tested the model's classification accuracy of the five standard

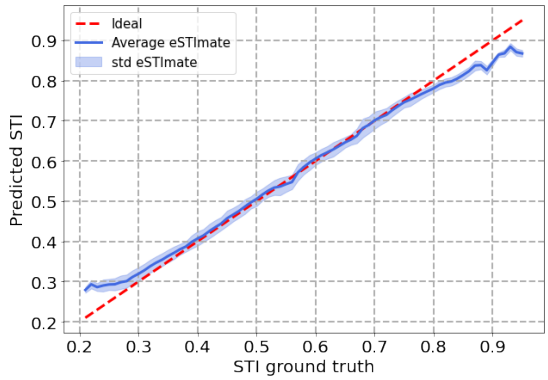

Figure 3: *Performance evaluation on the test set.*

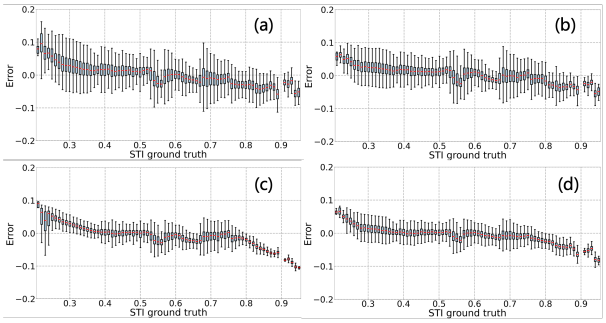

Figure 4: *Error in predicting STI for models in the ablation study with values closer to 0 indicating better performance. (a) Backbone. (b) FPN. (c) FPN with attention module. (d) FPN with attention module and automatic weight parameter learning.*

categories and achieved an accuracy of 92.75%, which further confirms the model's predictive capability.

4.4. Ablation study

On the basis of only using Resnet50 as backbone, we explored the influences of FPN structure, Multi-head Attention Block (MAB), and Automatic Weight Parameter Learning (AWPL) on STI estimation and noise-free spectrogram prediction respectively. The results are shown in Table 1 and Figure 4.

Before the introduction of AWPL, we initially set the weight parameters $\alpha = 1$ and $\beta = 0.5$ for FPN-based models. The results indicate that incorporating the FPN structure and auxiliary task led to a notable improvement on the STI prediction task, with MAE and RMSE enhanced by 14% and 16%, respectively.

After introducing the attention module, the STI prediction metrics remain relatively unchanged. Figure 4(c) shows that this is due to a relatively larger bias in the predictions and a reduction in variance, indicating an improvement in the model's generalization ability. In addition, the performance of the auxiliary speech enhancement task improved, suggesting that the preset parameters might have favored the auxiliary task.

With the introduction of AWPL, the performance of STI estimation is further improved, resulting in a 29% and 25% increase in MAE and RMSE, respectively, compared to the baseline. The model still maintains a small variance and further reduces the prediction bias for STI values less than 0.3 and greater than 0.8, as shown in Figure 4(d), achieving the best overall performance in the ablation experiment.

Table 1: *Ablation study on the proposed structure.* ***FPN****:Feature Pyramid Network.* ***MAB****: Multi-head Attention Block.* ***AWPL****:Automatic Weight Parameter Learning.*

Backbone	FPN	MAB	AWPL	MAE ↓	RMSE ↓	SSIM ↑
✓				0.023	0.028	-
✓	✓			0.019	0.024	0.714
✓	✓	✓		0.019	0.026	**0.719**
✓	✓	✓	✓	**0.016**	**0.021**	0.705

Table 2: *Blind STI estimation framework comparison.* ***MTL****:Multi-Task Learning of other room acoustic parameters.*

System	RMSE ↓	Validation Range
CNN [11]	0.037	0.65 - 0.96
TAE-CNN [12]	0.120	0.30 - 0.75
TAE-CNN-MTL [13]	0.040	0.35 - 0.80
CRNN-MTL [14]	0.033	0.38 - **1.00**
eSTImate	**0.021**	**0.21** - 0.95

4.5. Framework comparison

We conducted a comparison of the proposed eSTImate method with previous methods as described in Section 1. Table 2 presents the results, where RMSE is the primary evaluation metric. It should be noted that the STI distribution range differs in previous studies due to the use of different datasets and data generation methods, and the validation range is also listed in the comparison table to measure the coverage and overall capabilities of the models.

It is evident that previous STI framework cannot distinguish between the Poor and Bad levels in the 5-level standard STI classification, while our proposed model can effectively differentiate between all the 5 levels and achieves the lowest RMSE. The results demonstrate that the proposed eSTImate outperforms the state-of-the-art method by 36%, not only covering a wider range of STI predictions but also achieving smaller deviations.

5. Conclusions

In this study, we propose eSTImate, an end-to-end deep learning system for estimating the speech transmission index. Speech enhancement is introduced as an auxiliary task through a feature pyramid module to facilitate the network to learn better feature representations. To utilize the relationships between speech frames, a multi-head attention module is incorporated. Moreover, the technique of automatically learned weights is employed to better guide the gradient descent. We demonstrate the effectiveness of the proposed modules through ablation experiments, and compare our model with existing works to exhibit its state-of-the-art performance in providing accurate and consistent real-time STI estimation across diverse real-world scenarios.

6. References

[1] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (mos) revisited: Methods and applications, limitations and alternatives," *Multimedia Syst.*, vol. 22, no. 2, p. 213–227, mar 2016. [Online]. Available: https://doi.org/10.1007/s00530-014-0446-1

[2] 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," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, pp. 749–752 vol.2.

[3] a. w. rix, m. p. hollier, j. g. beerends, and a. p. hekstra, "pesq-the new itu standard for end-to-end speech quality assessment," *journal of the audio engineering society*, september 2000.

[4] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *The Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90–119, 1947.

[5] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *The Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 318–326, 1980. [Online]. Available: https://doi.org/10.1121/1.384464

[6] H. J. M. Steeneken and T. Houtgast, "Some applications of the Speech Transmission Index (STI) in auditoria," p. 11433, Feb. 1982.

[7] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[8] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387–3405, 2009.

[9] *Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index*, IEC IEC 60 268-16:2020, 2020.

[10] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Journal of the Acoustical Society of America*, vol. 54, pp. 557–557, 1973.

[11] P. Seetharaman, G. J. Mysore, P. Smaragdis, and B. Pardo, "Blind estimation of the speech transmission index for speech quality prediction," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 591–595.

[12] S. Duangpummet, J. Karnjana, W. Kongprawechnon, and M. Unoki, "A robust method for blindly estimating speech transmission index using convolutional neural network with temporal amplitude envelope," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 1208–1214.

[13] ——, "Blind estimation of speech transmission index and room acoustic parameters based on the extended model of room impulse response," *Applied Acoustics*, vol. 185, p. 108372, 2022.

[14] P. S. López, P. Callens, and M. Cernak, "A universal deep room acoustics estimator," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 356–360.

[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[16] M. R. Schroeder, "Modulation transfer functions: Definition and measurement," *Acta Acustica united with Acustica*, vol. 49, no. 3, pp. 179–182, 1981.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[19] L. Liebel and M. Körner, "Auxiliary tasks in multi-task learning," *arXiv preprint arXiv:1805.06334*, 2018.

[20] C. Warren, "Echothief impulse response library," www.echothief.com/downloads/, accessed: 2022-06-06.

[21] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[22] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[23] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1128–1132.

[24] ——, "Tut database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1128–1132.