COBEM 2021 26th International Congress of Mechanical Engineering
November 22-26, 2021, Virtual Congress, Brazil

COB-2021-1087
ASSESSMENT OF DEEP TRANSFER LEARNING FOR SPEECH TRANSMISSION INDEX PREDICTION

Eriberto Oliveira do Nascimento
Paulo Henrique Trombetta Zannin
Institution Laboratory of Environmental and Industrial Acoustics and Acoustic Comfort, Federal University of Paraná, Curitiba, 81530-000, Brazil
eriberto.on@gmail.com
paulo.zannin@gmail.com

Abstract. *Acoustic comfort in classrooms is a fundamental element that interferes with the teaching-learning dynamics. Speech intelligibility is measured using the Speech Transmission Index (STI). However, measuring STI is complex and requires expensive instrumentation. There are plenty of predictive models of STI that use the Reverberation Time as a regression variable. Still, these models do not consider the spectral effects of the room's impulsive response signal, as well as the spectral content of the background noise. This paper aimed to model STI through the one-deep dimensional convolutional neural network. The neural network had two inputs, simulated impulsive response via the Virtual Sources Method with 10,000 samples. The second input was the background noise with 564 samples. The training target was the predictive STI provided by IEC 60268-16. In addition, the Deep transfer learning method was employed. The transfer learning aims to optimize the previously trained neural network models with simulated data and transfer the generalization to in situ measurement data. Thus, in conclusion, the proposed model can be used as a support tool for diagnosing STI in classrooms.*

Keywords: *Deep Learning, Speech Transmission Index, Reverberation Time, Input Sensitivity*

1. INTRODUCTION

Noise as an environmental stressor is one of the major elements triggering diseases (GUO et al., 2017). Understanding the effects of noise depends on the factors that characterize it, such as the source, the transmission path, and the receiver. Each of these listed items can manifest in the most varied ways in everyday life, even without the awareness of the affected individuals, which is the case in school environments. Furthermore, noise pollution strongly interferes with the teaching-learning relationship. Therefore, it can cause illnesses in students and teachers related to prolonged exposure (BASNER et al., 2014).

In school environments, Connolly (2019) stated the main harmful effects of noise on the psychosocial development of children. Connolly (2019) also listed that such effects could hang on to the later stages of the child's development, reaching manifestations in later adulthood. So, there are reports of sleep disturbances, stress, mental fatigue, loss of concentration, communication problems, irritability, cognitive and learning disorders, and cardiovascular diseases (WEN et al., 2019).

Reverberation Time (RT) is one of the most consolidated acoustic descriptors and, therefore, widely used in architectural design. Recently, RT gained attention in automatic speech recognition systems due to the possibility of performing the convolution of the voice signal with an impulse response to produce advanced auralization (TENENBAUM et al., 2018).

However, according to Van Schoonhoven, Rhebergen, and Dreschler (2017), the reverberation time has several disadvantages when defining an interpretation regarding speech quality. Therefore, classical approaches in acoustics endeavored to characterize the optimal classroom design for speech based on RT, which is one of the most well-established descriptors employed throughout the 20th century. In this regard, other descriptors have been developed to ponder exogenous effects, such as Background Noise (BGN), to quantify the acoustic quality of rooms. As a result, the Speech Transmission Index (STI) was the leading objective descriptor designed to measure speech quality. Consequently, recognizing the influence of BGN on intelligibility was one of the driving forces for developing STI (Houtgast and Steeneken, 1980).

Subsequently, developing new diagnostic assessments for acoustic quality that are simultaneously fast and practical is paramount. Hence, this work aims to contribute to the development of these computational tools. The current available predictive models of STI prediction are vast, and they use different approaches to calculate RT. Although, as a result, there is no standardization regarding the octave bands applied as a regression variable to predict the STI (NOWOŚWIAT; OLECHOWSKA, 2016; LIU et al. 2020). Current models use reverberation time in octave bands as a regression variable (ESCOBAR; MORILLAS, 2015). On the other hand, the deep one-dimensional convolutional artificial neural networks

(1D CONV DNN) provide a model that uses the entire signal captured from the room's impulsive response. Additionally, it is possible to optimize the deep neural networks with the transfer learning heuristic to fill these gaps in the models.

2. STI PREDICTIVE MODEL

An analytical-predictive model for the STI is available in Annex J of IEC 60268-16 (IEC, 2011). The analytical model uses the Schroeder Equation to determine the modulation reduction factor as shown in Eq. 1.

$$m(f_m) \cong \left\{1 + \left(2\pi \frac{T}{13.8}\right)^2\right\}^{-1} \left[1 + 10^{-SNR/10}\right]^{-1} \quad (1)$$

with, SNR_k writing as

$$SNR_k = L_{op,k} - L_{BGN,k} \quad (2)$$

here, $m(f_m)$ are the modulation reduction factors, T is the reverberation time for each octave band, SNR is the signal-to-noise ratio, $L_{op,k}$ is the operational speech level of the source, and $L_{BGN,k}$ is the background noise level for the octave band k. After collecting the modulation factors, the STI is calculated according to the IEC 60268-16 (IEC, 2011) procedures. Finally, the $L_{op,k}$ is set according to the ANSI standard (ANSI, 2010), as shown in Table 1.

Table 1. Operational level (dB) in the receiver at one meter away from the source.

Level	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz	Leq
ANSI/ASA S12.60	#	57.2	59.8	53.5	48.8	43.8	38.6	59.5
IEC 60268-16	2.9	2.9	-0.8	-6.8	-12.8	-18.8	-18.8	0.0

The drawbacks of Eq. 1 rests in the difficulty of estimating the directivity factor of the source. According to Bistafa and Bradley (2000), the directivity factor compensates for the scattering effects when the distance between the source and the receiver is relatively large.

In this regard, Bistafa and Bradley (2000) proposed a model that mitigates the directionality factor. They argued that the directionality factor could be negligible depending on the distance since the reverberation energy levels supers the direct energy. For example, with an average 300 m^3 volume, the threshold distance is given as the direct ratio to the reflected sound energy density in classrooms. As a result, the threshold is the limit at which the direct energy is negligible, and the energy of the reverberant field reflections is the most significant. Consequently, the source directionality factor is inexpressive in the STI predictive models.

3. TRAINING DATASET CREATION

The STI measurements are time-consuming, e.g., STI takes up to 15 minutes for each measurement point. Thus, in medium-sized rooms, around 300 m^3, a uniform mesh of measuring points will take several hours. Besides, the IEC 60268-16 (IEC, 2011) requires a stationary background noise during measurement. The proposed solution to these problems was the creation of a database. This database was composed of rooms simulated impulse responses via the Image Method (ALLEN; BERKLEY, 1979) and the introduction of background noise through audio files from the Ko et al. (2017) database.

The reverberation time estimation was based on Allen and Berkley's (1979) code written in the FORTRAN language. This program was converted to C++ language and emulated in Python language. In total, 10,000 virtual rectangular rooms were designed with random dimensions as closed domain, with the coordinate system (x, y, z). The minimum volume was 120 $m^{3,}$ and the maximum was 1000 m^3.

The sound source was placed in the center of the rectangular room, 1.2 m above the base ground level, that is, with the coordinates at (xs, ys, 1.2 m), where xs and ys are the midpoints about the dimensions of room x and y, these also represent the positioning of the omnidirectional sound source. The simulated reverberation time ranged from 0.30 to 2.0 seconds. It was assigned a Gaussian distribution at the beginning of each simulation until obtaining 10000 impulse responses.

The signal samples that emulated background noise were derived from the environmental noise database (KO et al., 2017). The sampling frequency in these files has been set to 16 kHz. In total, 608 audio files. The sound pressure level (SPL) was normalized between zero and one Pascal in each audio sample. Then, the signal-to-noise ratio was calculated with the aid of an octave bandpass filter bank. Once obtained the time domain database for the impulse responses of the rooms and the background noise. These were post-processed using the spectral density, or power spectral density, given by,

$$S_{xx}(\omega) \int_{-\infty}^{\infty} |h(t)|^2 dt = \int_{-\infty}^{\infty} |H(\omega)|^2 \, d\omega \tag{3}$$

here, $h(t)$ is a temporal signal, and $H(\omega)$ is the Fourier transform of $h(t)$ which represents the room's impulsive response. The module of $|H(\omega)|$ is given by $H^*(\omega)H(\omega)$, where $H^*(\omega)$ is the complex conjugate of $H(\omega)$. The power spectral density (PSD) was estimated via the Welch periodogram method (WELCH, 1967). A frequency discretization with an increment of Δω of 1 rad/s was adopted, with frequencies ranging from 0 Hz to 8 kHz.

This processing was used to standardize the signal energy by bandwidth and set the same size for the input-output pair of artificial neural networks. Furthermore, in the formulation of indirect STI via Schroeder's Equation (SCHROEDER, 1981), the modulation transfer function can be interpreted as the complex Fourier transform of the impulse response power for a finite energy signal (HOUTGAST; STEENEKEN; PLOMP, 1980, BISTAFA; BRADLEY, 2000). Moreover, because of Parseval's Theorem, the signal energy invariance is observed, regardless of the time or frequency domain representation.

4. DEEP LEARNING

4.1 Training dataset

The training data were composed of tuples corresponding to two ordered inputs and one target value. The inputs were the spectral density of the impulse responses and the background noise. The output was the respective class of both input variables. In the training set, the input data were normalized by applying the transformation via z-score transformation.

The criterion for adopting the number of classes was based on the Just Noticeable Difference (JND) concept. The JND of the STI is 0.03, according to Bradley, Reich, e Norcross (1999). With the number of classes $k = 10$, the difference in the bin is 0.05, which is more significant than one JND, and less than 0.06, equivalent to twice the JND. Therefore, the network will generalize over the slightest possible difference without generating too many classes.

For each training sample, the one-hot encoding was used for class labeling. The "one-hot" coding infers that the deep neural network was applied as a classifier. Therefore, k = 1 to 10 represents the cardinality associated with the sample class i. Thus, k is the cardinality corresponding to the number of output classes given by $\{y_i\}_{j=1}^{k} \in \{0, +1\}$ used in the SoftMax output layer.

4.2 Model specification and training

Figure 1 shows the schematic representation of the topology of the DNN model. It was composed of a one-dimensional convolutional layer, followed by a fully densely connected layer and a SoftMax layer.

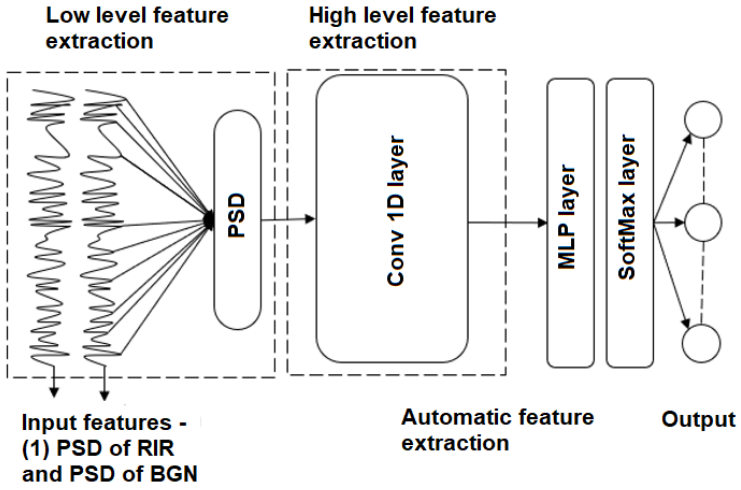

Figure 1. Automatic feature attribute extraction coupled with the Deep Neural Networks

The SoftMax output layer in Figure 1 compares the target and the predicted classes. It returns the conditional probability of the input sample belonging to the reference class, according to Eq. 4.

$$h_\theta(x^{(i)}) = \begin{bmatrix} p(y^{(i)} = 1|x^{(i)};\theta) \\ p(y^{(i)} = 2|x^{(i)};\theta) \\ \vdots \\ p(y^{(i)} = 1|x^{(i)};\theta) \end{bmatrix} = \frac{1}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}} \begin{bmatrix} e^{\theta_1^T x^{(i)}} \\ e^{\theta_l^T x^{(i)}} \\ \vdots \\ e^{\theta_k^T x^{(i)}} \end{bmatrix} \quad (4)$$

here $h_\theta(x^{(i)})$ is the probability distribution for obtaining the class $y^{(i)}$ given the input sample $x^{(i)}$. The predicted class is calculated as the maximum argument of the conditional probabilities, given by $\arg\max_{0 \le x \le k} h_\theta(x^{(i)})$

At the same time, the weights are updated in the densely connected layer through the traditional error backpropagation algorithm (Rumelhart, Hinton, and Williams, 1986). The weights were feedforwarded into the Conv1D layer according to the Zeng et al. (2014) model, given in Eq. 5,

$$w_{i,j} = \sum_{j}^{s} w_{j,i}^{l-1} \sigma(z_{i,j}) + w_{i,j}^{l-1} \quad (5)$$

here $w_{i,j}$ is the weight matrix, written in the indicial form. Therefore, the interaction of the feedforward step in the Conv1D layer is given by Eqs. 6 and 7 (ZENG et al., 2014).

$$w_{i,j} = w_{j,i}^{l-1} - \alpha \frac{\partial \mathcal{L}_i}{\partial w_{i,j}} \quad (6)$$

$$\frac{\partial \mathcal{L}_i}{\partial w_{i,j}} = \sum_{i=1}^{C} t_i \frac{\partial \mathcal{L}_i}{\partial w_{i,j}} \quad (7)$$

According to the ADAM algorithm, the loss function in Eq. 7 is optimized (CHOLLE et al., 2018). Furthermore, the ADAM algorithm performs stochastic optimization through first-order gradients (KINGMA; BA, 2014). Therefore, the relevant quality metrics for the classification problem were the Accuracy (ACC) and the F1-score.

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{C} \mathbf{1}(y_i) \frac{e^{w_k^T b_i}}{\sum_{j=1}^{C} e^{w_k^T b_i}} \quad (8)$$

Thus, the one-dimensional convolutional deep neural networks generated the latent space based on the input signal features. Latent spaces are generic denominations of tensor, vector, and subspaces synaptic weights distribution. After, the generated latent space was used in the deep transfer learning given in section 5.

5. DEEP TRANSFER LEARNING

Deep transfer learning aims to re-implement a DNN model fine-tuned for a specific dataset onto another. The novel dataset must possess a similar topological structure or manifest a relatively high likelihood ratio in the probability space. Statistically, this problem occurs in the relations between the source and target domains. Their affinity can be understood by learning rules based on projections (ZHUANG et al., 2020). The Maximum Mean Discrepancy (MMD) showed in Eq 9 was one of the first attempts to semi-empirically quantify a distribution-distance metric to express the statistical similarity of samples from the training subspace.

$$MMD(X^S, X^T) = \left\| \frac{1}{n^S} \sum_{i=1}^{n^S} \Phi(x_i^S) - \frac{1}{n^T} \sum_{j=1}^{n^T} \Phi(x_i^T) - \right\|_{\mathcal{H}}^2 \quad (9)$$

here X^S and X^T are the number of samples in the source and target domains, respectively. The Φ is a function that transforms the vector space of the samples to a dimension in the latent space. The $\mathcal{H}$ space corresponds to the Reproducing Kernel Hilbert Space (RKHS),

The Variational Autoencoder (VAE) acts as a proxy to the Φ function shown in Eq. 9. The VAE produces a meta-representation via the functional device $f(z;w)$: $\mathbb{F} \rightarrow \mathbb{X}$, $z \in \mathbb{F}$ indicates the latent space, x $\in$ $\mathbb{X}$ corresponds to the feature space, w is a vector that parameterizes function f. The autoencoder has three elements, the encoder, the code, and the decoder. The input-output mapping is given by, x→x̃, where $\tilde{x}$ is an estimator of the input x, in general terms, $f(z;w)$

Produces an identity operator, by the dimensional strangulation of x, via the latent space z. The estimate is given by, $\tilde{x} = f_d(f_e(z; w_e); w_d)$, f_d is the decoder and f_e is the encoder (LE; PATTERSON; WHITE, 2018).

The VAE focuses on the nonlinear dimensionality reduction of the feature space to z − space by adding one more layer responsible for the statistical characterization of sample x. Therefore, the dimensionality reduction was performed using the Bayesian stochastic gradient method. The objective function must have at least the first-order derivative defined, i.e., a Class 1 function. The attribute space is called latent space, given by, $z_i = \mu_i + \sigma_i \hat{z}_i$, where $\hat{z}_i \sim \mathcal{N}(0,1)$ represents a Gaussian probability density function. The Gaussian one, $\mathcal{N}(0,1)$ has a null mean and a unit standard deviation with the entire sample space.

Finally, the latent space generated on the 1D CONV DNN model trained in section 4 was set as the input trained data from the VAE. Subsequently, the VAE network produces a reduced representation of x applied in Eq. 9 to perform the MMD, generating the (KINGMA; WELLING, 2013).

6. RESULTS AND DISCUSSION

Through the image source method, 10,000 acoustic virtual rooms were created. Figure 1 shows an example of the acoustic model and its respective Room Impulse Response (RIR) captured at the receiver position.

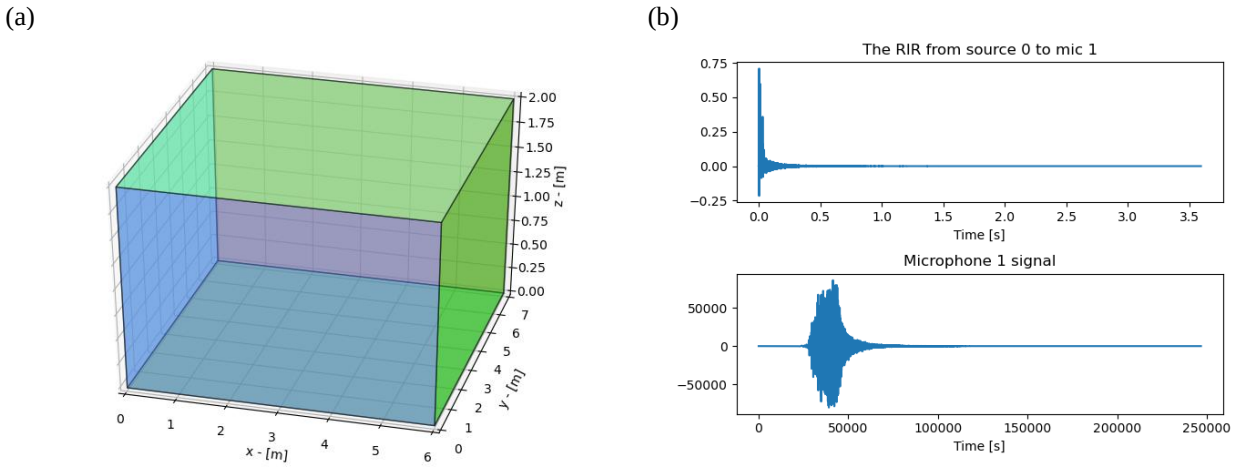

Figure 2. (a) Virtual room generated for the Image Source Method, (b) Example of the RIR captured at a random position inside the room

Figure 3 shows the broadband distribution of the background noise level extracted from the audio samples from the database of [3]. The background noise was transformed to the one-third octave band filter, and through Eq. 1 and Eq. 2, the analytical values of STI were obtained.

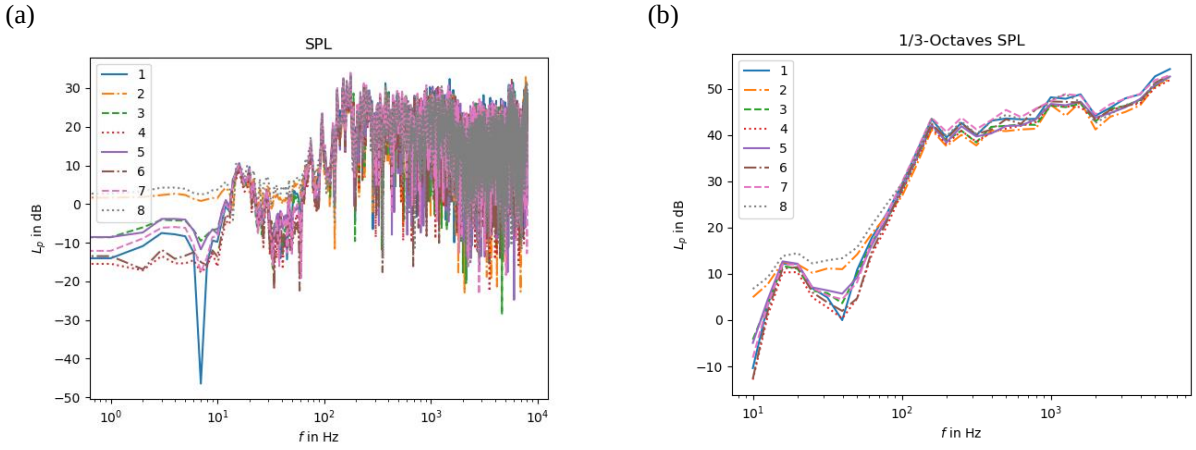

Figure 3. (a) Broadband acoustical signal extracted from the environmental noise database from Ko et al. (2017), (b) Acoustic signal at 1/3 octave band filter

Figure 4 shows the STI histogram to the (n = 10000) samples and the (k = 10) classes. These results agree with the literature stating that the STI values are concentrated in 0.2 to 0.6.

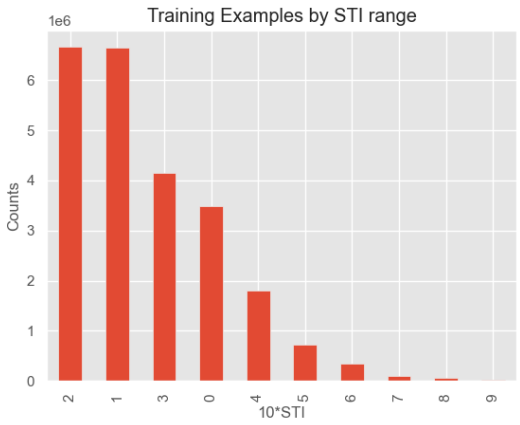

Figure 4. Histogram of training sample for STI.

The power density spectrogram of the RIR and the broadband background noise signals, as shown in Fig. 5, was fed to the DNN as the sample input and their respective STI value as the target values.

(a) (b)

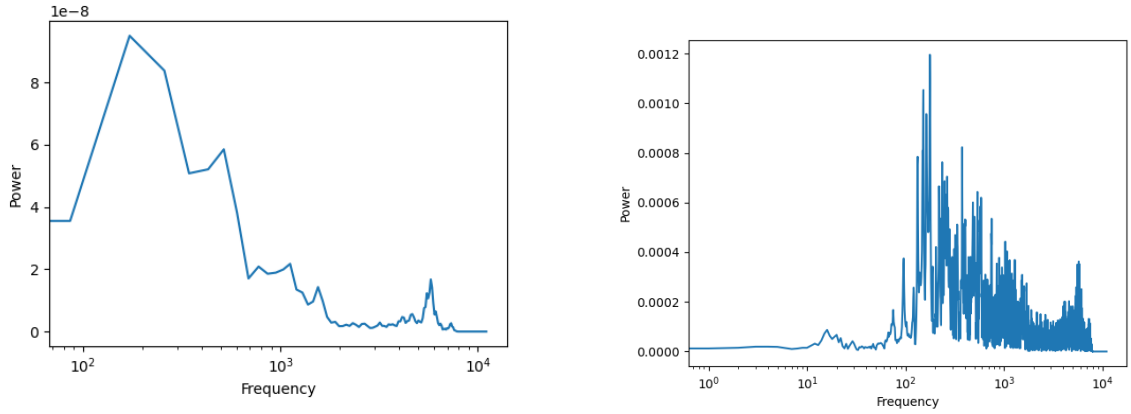

Figure 5. Samples power spectral density (PSD) of: (a) simulated RIR (b) Background noise signal

After the creation of the dataset was possible to proceed to the deep neural network training. Figure 5 shows the overall model accuracy, which went above 0.9 for the training data set.

(a) (b)

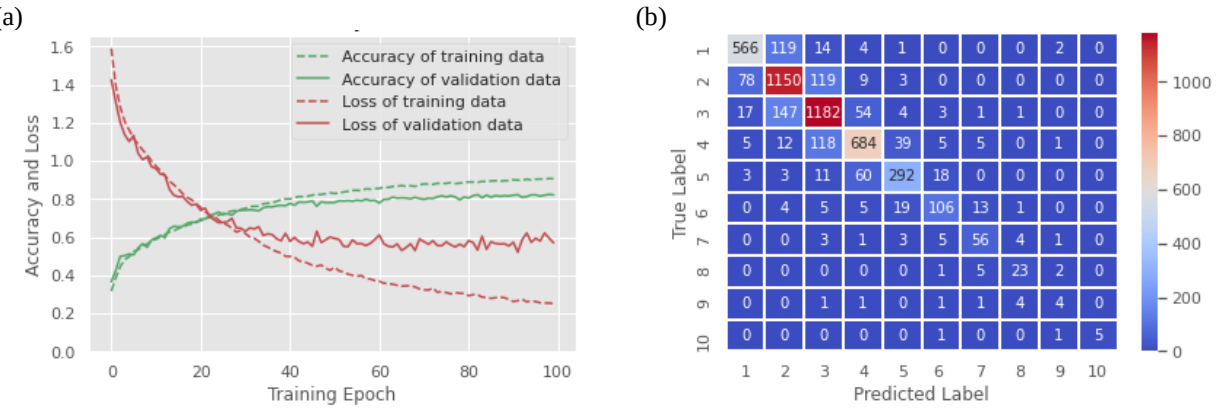

Figure 6. (a) DNN accuracy and loss (b) Confusion Matrix

Figure 6(b) shows the confusion matrix for the entire model; it was possible to assert that the model was calibrated and that the F1 score of 0.85 and the accuracy score were above 0.9, which indicates an excellent generalization to the model. Regarding transfer learning, Figure 7 shows the respective values for the classes assigned into the latent space with a projection on turn the two first dimensions through the VAE.

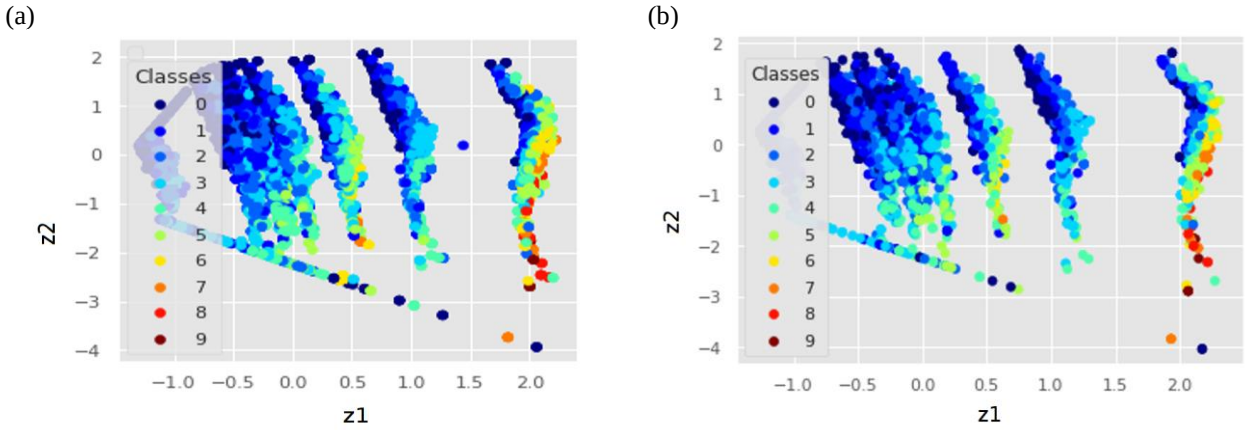

Figure 7. VAE class distribution projection over the latent space z1-z2 dimensions (a) Training dataset (b) Test dataset

In the training process, it was used two datasets, the training and the test dataset. The test was not presented during the training of the deep neural network. As a result, the class distribution was uniform for both datasets. Finally, the autoencoder architecture constructed a simplified representation of the class in latent space projected onto a bidimensional space. As a result, Figure 6(a) shows an example of the class k = 1 for the training data set, in Figure 6(b) for the same class but the unseen data on the training test dataset.

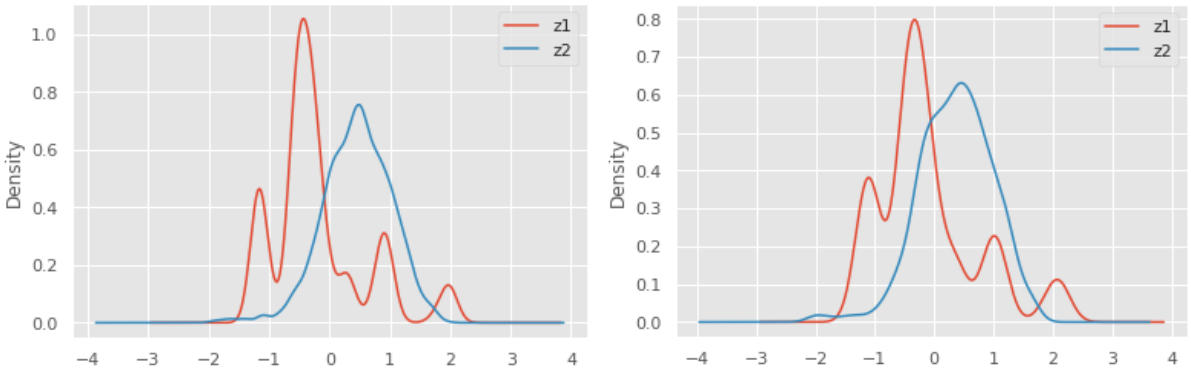

Figure 8. Simplified latent coordinates space (z1,z2) from the MMD metric (a) Class k = 1 for the training dataset (b) Predicted class k = 1 for the test dataset, which was unseen during the DNN training

Comparing z1 and z2 latent coordinates from the MMD, Figures 8(a) and 8(b) show a significant concordance level. For the z1-coordinate, the linear squared correlation of 0.95 and z2-coordinates of 0.95. Therefore, using Eq. 9, the predicted class was assigned correctly because the metric distance MMD was near zero since the higher values of linear correlations. The same result was found in the other classes.

7. CONCLUSIONS

This work aimed to evaluate a methodology for calculating and predicting the Speech Transmission Index in classrooms. To this end, a method based on the transfer learning applied in deep learning models was employed. As a result, the goodness of fit and the prediction for data not submitted to the training generate good results. Furthermore, it was found that using transfer learning could predict the STI based only on the reverberation time measurements using their time-domain version for both the room impulsive response and background noise value. Thus, it is concluded that the proposed model can be used as a support tool for a diagnostic evaluation of STI in classrooms, using solely the instrumentation used for RT measurements. So, the results showed that the DNN model predicted the STI with great accuracy.

Besides, new methods for measuring STI with non-stationary background noise are being developed. Therefore, this work showed the potential usage of Deep Transfer Learning to predict the STI in such situations. Consequently, the Maximum Mean Discrepancy (MMD) coupled with the Variational Autoencoder (VAE) with a stack approach can be applied in other fields of study, employing deep transfer learning heuristics.

8. ACKNOWLEDGEMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

9. REFERENCES

ALLEN, J. B.; BERKLEY, D. A. Image method for efficiently simulating small-room acoustics. The Journal of the Acoustical Society of America, v. 65, n. 4, p. 943-950, 1979.

AMERICAN NATIONAL STANDARD INSTITUTE (ANSI). ANSI/ASA S12.60-2010/PART 1: Acoustical Performance criteria, Design Requirements, and guidelines for Schools, Part 1: Permanent Schools. Acoustic Society of America, 2010.

BASNER, M. et al. Auditory and non-auditory effects of noise on health. The lancet, v. 383, n. 9925, p. 1325-1332, 2014.

BISTAFA, S. R.; BRADLEY, J. S. Reverberation time and maximum background-noise level for classrooms from a comparative study of speech intelligibility metrics. The Journal of the Acoustical Society of America, v. 107, n. 2, p. 861-875, 2000.

BRADLEY, J. S.; REICH, R.; NORCROSS, S. G. A just noticeable difference in C50 for speech. Applied Acoustics, v. 58, n. 2, p. 99-108, 1999.

CHOLLET, F. al. Keras: The python deep learning library. ascl, p. ascl: 1806.022, 2018.

CONNOLLY, D. et al. The effects of classroom noise on the reading comprehension of adolescents. The journal of the Acoustical Society of America, v. 145, n. 1, p. 372-381, 2019.

ESCOBAR, V. G; MORILLAS, J. M B. Analysis of intelligibility and reverberation time recommendations in educational rooms. Applied Acoustics, v. 96, p. 1-10, 2015.

GUO, L. et al. Effects of environmental noise exposure on DNA methylation in the brain and metabolic health. Environmental research, v. 153, p. 73-82, 2017.

HOUTGAST, T.; STEENEKEN, H. JM; PLOMP, R. Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics. Acta Acustica united with Acustica, v. 46, n. 1, p. 60-72, 1980.

INTERNATIONAL ELECTROTECHNICAL COMMISSION (IEC). IEC 60268-16: Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index. Switzerland, 2011.

KINGMA, D. P.; BA, Jimmy. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

KINGMA, D. P.; WELLING, M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

KO, T. et al. A study on data augmentation of reverberant speech for robust speech recognition. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017. p. 5220-5224.

LE, L.; PATTERSON, A.; WHITE, M. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. In: Advances in Neural Information Processing Systems. 2018. p. 107-117.

LIU, H. et al. The speech intelligibility and applicability of the speech transmission index in large spaces. Applied Acoustics, v. 167, p. 107400, 2020.

NOWOŚWIAT, A.; OLECHOWSKA, M. Fast estimation of speech transmission index using the reverberation time. Applied Acoustics, v. 102, p. 55-61, 2016.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, Ronald J. Learning representations by back-propagating errors. Nature, v. 323, n. 6088, p. 533-536, 1986.

SCHROEDER, M. R. Modulation transfer functions: Definition and measurement. Acta Acustica united with Acustica, v. 49, n. 3, p. 179-182, 1981.

TENENBAUM, R. A. et al. Auralization generated by modeling HRIRs with artificial neural networks and its validation using articulation tests. Applied Acoustics, v. 130, p. 260-269, 2018.

VAN SCHOONHOVEN, J.; RHEBERGEN, K. S.; DRESCHLER, W. A. Towards measuring the Speech Transmission Index in fluctuating noise: Accuracy and limitations. The Journal of the Acoustical Society of America, v. 141, n. 2, p. 818-827, 2017.

WELCH, P. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Transactions on audio and electroacoustics, v. 15, n. 2, p. 70-73, 1967.

WEN, X. et al. Impacts of traffic noise on roadside secondary schools in a prototype large Chinese city. Applied Acoustics, v. 151, p. 153-163, 2019.

ZENG, M. et al. Convolutional neural networks for human activity recognition using mobile sensors. In: 6th International Conference on Mobile Computing, Applications and Services. IEEE, 2014. p. 197-205.

ZHUANG, F. et al. A comprehensive survey on transfer learning. Proceedings of the IEEE, 2020.

10. RESPONSIBILITY NOTICE

The authors are the only ones responsible for the printed material included in this paper.