STOI-Net: A Deep Learning based Non-Intrusive Speech Intelligibility Assessment Model

Ryandhimas E. Zezario*†, Szu-Wei Fu†, Chiou-Shann Fuh*, Yu Tsao†, Hsin-Min Wang‡
*Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
E-mail: fuh@csie.ntu.edu.tw
†Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
E-mail: {ryandhimas, jasonfu, yu.tsao}@citi.sinica.edu.tw
‡Institute of Information Science, Academia Sinica, Taipei, Taiwan
Email: whm@iis.sinica.edu.tw

***Abstract*— The calculation of most objective speech intelligibility assessment metrics requires clean speech as a reference. Such a requirement may limit the applicability of these metrics in real-world scenarios. To overcome this limitation, we propose a deep learning-based non-intrusive speech intelligibility assessment model, namely STOI-Net. The input and output of STOI-Net are speech spectral features and predicted STOI scores, respectively. The model is formed by the combination of a convolutional neural network and bidirectional long short-term memory (CNN-BLSTM) architecture with a multiplicative attention mechanism. Experimental results show that the STOI score estimated by STOI-Net has a good correlation with the actual STOI score when tested with noisy and enhanced speech utterances. The correlation values are 0.97 and 0.83, respectively, for the seen test condition (the test speakers and noise types are involved in the training set) and the unseen test condition (the test speakers and noise types are not involved in the training set). The results confirm the capability of STOI-Net to accurately predict the STOI scores without referring to clean speech.**

I. INTRODUCTION

For many speech-related applications, such as assistive oral communication devices [1-5] and telecommunications [6-8], and speech-related tasks, such as speech coding [9, 10], voice conversion [11, 12], speech separation [13, 14], and speech enhancement [15-18], speech intelligibility plays a crucial role in determining the performance of processed speech signals. An intuitive method to measure speech intelligibility is to conduct a human listening test. By playing test samples to subjects, the intelligibility scores can be calculated by the ratio of the number of accurately recognized words to the total number of words in the played speech samples. To make an accurate and unbiased evaluation of speech intelligibility, it is necessary to recruit as many subjects as possible, and each subject must listen to a large number of test utterances covering diverse conditions. In general, this may be prohibitive and may not be feasible. Therefore, several approaches have been proposed to estimate speech intelligibility as surrogates for the human listening test [19-23].

The articulation index (AI) [19] and speech intelligibility index (SII) [20] are two well-known objective speech intelligibility predictors; both metrics have been widely used to measure speech intelligibility in various speech-related applications. Based on the design of SII, extended SII (ESII) [24] and coherence SII (CSII) [25] were derived to attain better intelligibility measurements. The speech transmission index (STI) [21, 22] extends the range of distortion to convolutive noise (e.g., reverberant speech and effects of room acoustics) by considering the depth of temporal signal modulation compared to the clean, undistorted reference signal. Recently, short-time objective intelligibility (STOI) [23] has been proposed. Its calculation consists of five major steps: (1) silent frame removal, (2) short-time Fourier transform (STFT), (3) one-third octave band analysis, (4) normalization and clipping, and (5) intelligibility measurement. In terms of predictive ability, STOI has shown a notable improvement in intelligibility scoring in several domains [26-28] over the previous methods, and has therefore been widely used as a standardized evaluation metric for many speech-related tasks. A notable limitation of STOI, however, is the requirement for clean speech as a reference, which may not always be accessible, especially during online operations. Several extensions to address this issue have been developed, such as non-intrusive STOI [29].

Although the traditional signal processing-based intelligibility assessment metrics have shown satisfactory measurement results and have been widely adopted as assessment tools for various speech-related tasks, their applicability is still limited because of the following two factors. (1) The generalization of these metrics to new conditions still has room for improvement. Particularly, the assessment performance may degrade while operating under new and unseen conditions. Even if some training data for the new conditions are available, the assessment metrics cannot be adapted to the new data. (2) The compatibility of these metrics to speech processing systems, which are usually built based on deep neural networks in recent years, is restricted. More specifically, these traditional metrics cannot be readily integrated with speech-related systems (such as noise reduction and speech separation) to jointly optimize the overall performance. Due to the above limitations, it is crucial to determine an effective intelligibility assessment metric that can be continuously learned to adapt to new test conditions and can be easily combined with (learning-based) speech processing systems.

In our previous work [30], we had developed a neural network-based non-intrusive quality assessment model, namely

Quality-Net, to estimate the perceptual evaluation of speech quality (PESQ) score [31]. Quality-Net has shown remarkable performance in evaluating noisy and processed speech without the need for clean speech as a reference. As Quality-Net is based on deep neural network architecture, its prediction ability for new environments can be improved by adapting to the corresponding new data. Moreover, several studies have combined Quality-Net with speech-related systems to jointly optimize the overall performance [32, 33]. Along with this research direction, this study investigates and develops STOI-Net, a deep neural network model that can accurately predict STOI scores without the need for clean speech as a reference.

The proposed STOI-Net is formed by the combination of a convolutional neural network and bidirectional long short-term memory (CNN-BLSTM) architecture with a multiplicative attention mechanism. The CNN is used to extract informative features from the input data, and the BLSTM is used to model time-variant characteristics. The attention mechanism aims to boost the performance by focusing on important regions while calculating intelligibility scores. Experimental results reveal that the predicted scores yielded by STOI-Net have rather high correlation with the ground-truth STOI scores when tested in both seen (the test speakers and noise types are involved in the training) and unseen (the test speakers and noise types are not involved in the training) conditions. It may be noted that the STOI calculation requires clean speech as a reference, whereas STOI-Net does not. The results confirm the decent ability of STOI-Net to accurately predict STOI scores (evaluation of speech intelligibility) without the need for clean speech as a reference.

The remainder of this paper is organized as follows. Section II introduces the proposed STOI-Net. Section III describes the experimental setup and results. Finally, conclusions and future work are presented in Section IV.

II. STOI-NET

In this section, we introduce the model architecture and training objective of the proposed STOI-Net.

A. Architecture

Fig. 1 shows the overall architecture of STOI-Net, which consists of several stages. The input to STOI-Net is a sequence of spectral features of noisy/processed speech, and the output is the predicted STOI score. In STOI-Net, the CNN module has 12 convolutional layers, which are used to obtain informative features from the spectral features. Next, the BLSTM module is used to further model the temporal characteristics of the extracted features from the CNN. The attention mechanism is used to identify and weight the important regions in the input features. In our implementation, multiplicative attention is used to form the attention layer because of its high efficiency and satisfactory performance [34]. Next, a fully connected layer is used to map the frame-wise features into frame-wise scores. Finally, based on these estimated frame-level scores, a global average operation is applied to calculate the final predicted STOI score.

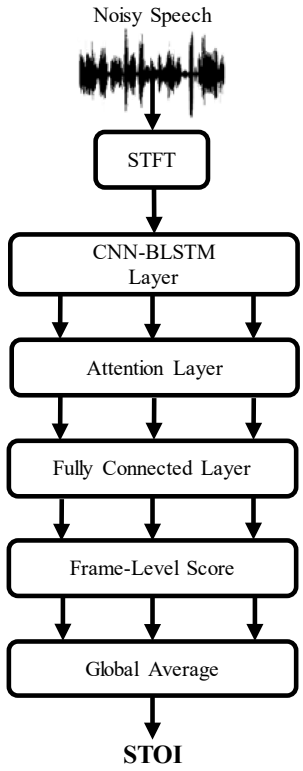

Fig. 1: Architecture of the STOI-Net model.

B. Objective Function

STOI-Net aims to estimate an utterance-level intelligibility score. However, because a speech utterance may contain non-stationary noises or distortions in different regions (segments of frames), directly assigning an utterance-level score to train STOI-Net may not be a suitable approach. Therefore, we prepare the frame-level scores to train STOI-Net. With the frame-level scores, the objective function can be derived as follows.

$$O = \frac{1}{N}\sum_{n=1}^{N}[(I_n - \hat{I}_n)^2 + \frac{1}{L(U_n)}\sum_{t=1}^{L(U_n)}\alpha(I_n)(I_n - \hat{\imath}_{n,t})^2] \quad (1)$$

where I_n and $\hat{I}_n$ are the true and predicted utterance-level STOI scores, respectively; N denotes the total number of training utterances; $L(u_n)$ denotes the number of frames in the n-th utterance; $\hat{\imath}_{n,t}$ is the predicted frame-level STOI score of the t-th frame of the n-th utterance; and $\alpha(I_n)$ denotes the weighting scale, which is determined by the attention mechanism. It can be seen that the first term estimates the accuracy of utterance-level scoring, and the second term estimates the accuracy of frame-level scoring. We believe that with the objective function in Eq. (1), STOI-Net can be trained to model the STOI metric locally and globally.

III. EXPERIMENTS

A. Experimental Setup

The Wall Street Journal (WSJ) dataset [35] was used to prepare the training and test sets in this study. The training set of the WSJ dataset contained 37,416 utterances, while the test set of the WSJ dataset contained 330 utterances, all recorded at a sampling rate of 16 kHz. We polluted the training utterances with 100 types of noise [36], covering both stationary and non-stationary noise types, with 31 different SNR levels ranging from -10 to 20 dB with an interval of 1 dB. A pretrained speech enhancement (SE) model was used to process the noisy utterances to obtain enhanced utterances. The SE model was formed by a BLSTM model with two bidirectional hidden layers, each consisting of 300 neurons. We randomly selected 15,000 noisy utterances, 15,000 enhanced utterances, and 1,500 original clean utterances to form the training set for the proposed STOI-Net model. The 1,500 clean utterances were used to let the STOI-Net learn the highest STOI score (i.e., 1.0).

We prepared two test sets to evaluate the STOI-Net model: the seen and unseen test sets. For the seen test set, we randomly selected 2,350 noisy utterances, 2,350 enhanced utterances, and 300 clean utterances from the remaining utterances in the above large training set. Thus, the seen test set contained a total of 5,000 utterances. It may be noted that the speakers and noise types overlap with those in the training utterances, but the utterances are different. For the unseen test set, we randomly selected 300 utterances from the test set of the WSJ dataset. For this test set, the speakers and noise types were different from those in the training utterances. We used four other noise types (car, pink, street, and babble) and contaminated the speech utterances at 6 SNR levels (-10, -5, 0, 5, 10, and 15 dB). Finally, we randomly selected 2,350 noisy utterances, 2,350 enhanced utterances, together with the 300 clean utterances, to form the unseen test set (a total of 5,000 utterances).

Each utterance in the training and testing sets was converted into a 257-dimensional spectrogram by applying a 512-point STFT with a Hamming window of 32 ms and a hop of 16 ms, which was used as the input for the STOI-Net. Three evaluation metrics, namely mean square error (MSE), linear correlation coefficient (LCC), and Spearman's rank correlation coefficient (SRCC), were used to evaluate the predicted STOI scores.

B. Effect of Model Architecture

First, we analyzed the prediction capability of STOI-Net with different model architectures. In a previous work [30], BLSTM has shown its advantage in modeling time-variant speech patterns. Therefore, we used the BLSTM model as our baseline system in this study. The proposed STOI-Net was formed by a CNN-BLSTM architecture, which included 12 convolutional layers, each consisting of four channels {16, 32, 64, and 128}, a one-layered BLSTM (with 128 nodes), and a fully connected layer (with 128 neurons).

Table 1 presents the LCC, SRCC, and MSE results of the BLSTM and CNN-BLSTM models under the seen test condition. Higher LCC and SRCC scores denote better results, while lower MSE scores denote better results. From Table 1, it can be seen that CNN-BLSTM outperforms BLSTM consistently, with higher LCC and SRCC scores and a lower MSE score. Table 2 presents the LCC, SRCC, and MSE results of the BLSTM and CNN-BLSTM models under the unseen test condition. In this table, the same trend as in Table 1 can be seen; in other words, compared with the BLSTM baseline, CNN-BLSTM can yield higher LCC and SRCC scores and a lower MSE score. In the following discussion, we will further evaluate the STOI-Net that is developed based on the CNN-BLSTM architecture.

Table 1. LCC, SRCC, and MSE results of BLSTM and CNN-BLSTM under the seen test condition.

	LCC	**SRCC**	**MSE**
BLSTM [30]	0.923	0.928	0.005
CNN-BLSTM	**0.964**	**0.962**	**0.002**

Table 2. LCC, SRCC, and MSE results of BLSTM and CNN-BLSTM under the unseen test condition.

	LCC	**SRCC**	**MSE**
BLSTM [30]	0.764	0.784	0.029
CNN-BLSTM	**0.789**	**0.797**	**0.016**

C. Effect of Attention Mechanism

From the previous experiment, we have confirmed that compared with BLSTM, CNN-BLSTM can achieve better prediction performance for STOI-Net. In this set of experiments, we aimed to further improve the prediction performance by adding a multiplication attention mechanism to STOI-Net; the system is termed CNN-BLSTM$_{ATT}$. From Table 3, it can be seen that under the seen test condition, CNN-BLSTM$_{ATT}$ can achieve higher LCC and SRCC scores compared to CNN-BLSTM. It should also be noted that the MSE score of CNN-BLSTM$_{ATT}$ is lower than that of CNN-BLSTM, but the improvement is small. Table 4 presents the results under the unseen test conditions. The results again show that with the attention mechanism, CNN-BLSTM$_{ATT}$ always produces better LCC, SRCC, and MSE scores, as compared to CNN-BLSTM without the attention mechanism.

Table 3. LCC, SRCC, and MSE results of CNN-BLSTM and CNN-BLSTM-ATT under the seen test condition.

	LCC	**SRCC**	**MSE**
CNN-BLSTM	0.964	0.962	0.002
CNN-BLSTM$_{ATT}$	**0.970**	**0.968**	**0.001**

Table 4. LCC, SRCC, and MSE results of CNN-BLSTM and CNN-BLSTM-ATT under the unseen test condition.

	LCC	**SRCC**	**MSE**
CNN-BLSTM	0.789	0.797	0.016
CNN-BLSTM$_{ATT}$	**0.827**	**0.815**	**0.015**

To study the reasons for the performance improvement provided by the attention mechanism, we further analyzed the CNN-BLSTM and CNN-BLSTM$_{ATT}$ models by visualizing the

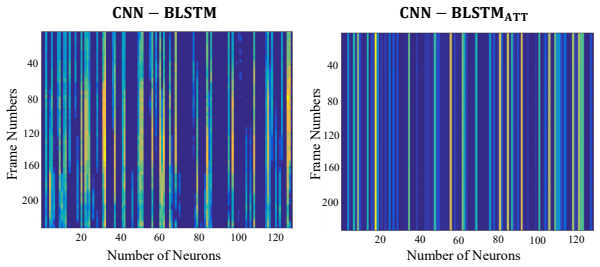

Fig. 2: Representations of a speech utterance at the hidden layers of CNN-BLSTM and CNN-BLSTM$_{ATT}$ models

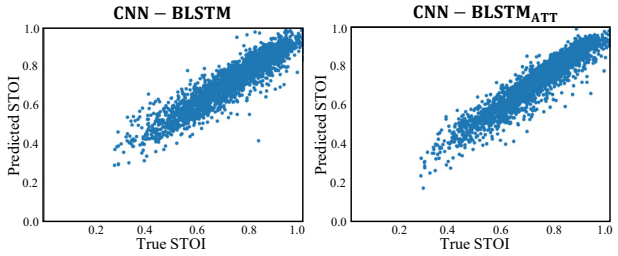

Fig. 3: Scatter plots of speech intelligibility assessment by STOI-Net.

hidden layers. As shown in Fig. 2, the representations of the hidden layers of CNN-BLSTM and CNN-BLSTM$_{ATT}$ show different patterns, confirming that the attention layer provides additional weights to specific regions. In addition, the scatter plots of speech intelligibility assessment by STOI-Net are shown in Fig. 3. We compared the scatter plots of the predicted scores generated by CNN-BLSTM and CNN-BLSTM$_{ATT}$. From the figure, it is clear that CNN-BLSTM$_{ATT}$ can predict STOI scores more accurately than CNN-BLSTM. This further shows that the proposed STOI-Net using CNN-BLSTM and the attention mechanism can achieve higher correlation performance, as compared to STOI-Net using CNN-BLSTM without an attention mechanism.

IV. Conclusions

In this study, we proposed STOI-Net, a deep neural network-based non-intrusive speech intelligibility assessment model. We aimed to use the STOI-Net as a surrogate to the traditional STOI evaluation metric. Experimental results first confirmed that the predicted scores of STOI-Net have a good correlation with the ground-truth STOI scores. Then, we confirmed the advantages of using CNN-BLSTM over BLSTM to form the STOI-Net model architecture under both seen and unseen test conditions. Finally, we confirmed the effectiveness of the attention mechanism, which can further improve the prediction performance. In the future, we will further evaluate the generalization ability by testing the STOI-Net prediction results under completely different test conditions from those in the training set. We will also explore the integration of STOI-Net into numerous speech-related applications to directly improve the performance of the target tasks.

References

[1] Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, "A Deep Denoising Autoencoder Approach to Improving the Intelligibility of Vocoded Speech in Cochlear Implant Simulation," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1568-1578, 2017.

[2] S. C. Peng, L. J. Spencer, and Tomblin, J. B. "Speech intelligibility of pediatric cochlear implant recipients with 7 years of device experience," *Journal of Speech, Language, and Hearing Research,* vol. 47, pp. 1227– 1237, 2004.

[3] M. Anshori, "Speech intelligibility and auditory perception of pre-school children with Hearing Aid, cochlear implant and Typical Hearing" *Otology*, vol. 15, pp. 62-66, 2020.

[4] R.-Y. Tseng, T.-W. Wang, S.-W. Fu, C.-Y. Lee, and Y. Tsao, "A Study of Joint Effect on Denoising Techniques and Visual Cues to Improve Speech Intelligibility in Cochlear Implant Simulation," *to appear in IEEE Transactions on Cognitive and Developmental Systems.*

[5] S.-W. Fu, P.-C. Li, Y.-H. Lai, C.-C. Yang, L.-C. Hsieh, and Y. Tsao, "Joint Dictionary Learning-based Non-Negative Matrix Factorization for Voice Conversion to Improve Speech Intelligibility After Oral Surgery," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 11, pp. 2584 - 2594, 2017.

[6] R. M. Ullmann, "Can you hear me now? Automatic assessment of background noise intrusiveness and speech intelligibility in telecommunications," *Ph.D. dissertation, Department Electrical Engineering, EPFL., Lausanne, Swiss*, 2016.

[7] S. Brachmanski, "Automation of Subjective Measurements of Speech Intelligibility in Analogue Telecommunication Channels," *Archives of Acoustic*s, Vol. 33, No. 3, 2008, pp. 341–350.

[8] 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, 2016.

[9] R. Martin and R. V. Cox, "New speech enhancement techniques for low bit rate speech coding," *in Proc. IEEE Workshop on Speech Coding,* pp. 165–167, 1999.

[10] Z. Zhao, H. Liu, and T. Fingscheidt, "Convolutional neural networks to enhance coded speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 663–678, 2019.

[11] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," *in Proc. IEEE ICASSP*, pp. 4869–4873, 2015.

[12] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with wavenet-based waveform generation," *in Proc. Interspeech*, pp. 1138–1142, 2017.

[13] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," *in Proc. IEEE ICASSP*, pp. 708–712, 2015.

[14] Z.-Q. Wang and D. Wang, "Recurrent deep stacking networks for supervised speech separation," *in Proc. IEEE ICASSP*, pp. 71–75, 2017.

[15] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 7–9, 2015.

[16] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," *in Proc. ICASSP*, pp. 5220–5224, 2016.

[17] R. E. Zezario, T. Hussain, X. Lu, H.-M. Wang, and Y. Tsao, "Self-supervised denoising autoencoder with linear regression

decoder for speech enhancement," *in Proc. ICASSP*, 2020, pp. 6669–6673.
[18] P. C. Loizou, "Speech enhancement: theory and practice," *CRC Press*, 2007.
[19] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds." Journal of the Acoustical Society of America, vol. 19, no. 1, pp. 90-119, 1947.
[20] ANSI Std. S3.5 1997, "Methods for calculation of the speech intelligibility index" 1997.
[21] T. Houtgast and H. 1. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, vol. 25, no. 6, pp. 355-367, 1971.
[22] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *Journal of the Acoustical Society of America*, vol. 67, no. 1, pp.318-326, 1980.
[23] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.
[24] K. Rhebergen, N. Versfeld, and W. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *Journal of the Acoustical Society of America,* vol. 120, pp. 3988–3997, 2006.
[25] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2224–2237, 2005.
[26] S. Jorgensen, J. Cubick, and T. Dau, "Speech intelligibility evaluation for mobile phones," *Acustica*, vol. 101, pp. 1016-1025, 2015.
[27] C. Yu, R. E. Zezario, J. Sherman, Y.-Y. Hsieh, X. Lu, H.-M. Wang, and Y. Tsao, "Speech enhancement based on denoising autoencoder with multi-branched encoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2756 – 2769, 2020.
[28] T. H. Falk, V. Parsa, 1. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114-124, 2015.
[29] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.
[30] S.-W. Fu., Y. Tsao, H.-T. Hwang, and H.-W. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," *in Proc. Interspeech*, pp. 1873–1877, 2018.
[31] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *Int. Telecommun. Union, T Recommendation*, no. 862, pp. 708-712, 2001.
[32] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," *in Proc. ICML*, 2019
[33] R. E. Zezario, S.-W. Fu, X. Lu, H.-M. Wang, and Y. Tsao, "Specialized Speech Enhancement Model Selection Based on Learned Non-Intrusive Quality Assessment Metric," *in Proc. Interspeech*, pp. 3167-3172, 2019.
[34] A.Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *In Proc. NIPS*, pp 6000–6010, 2017
[35] D. Paul and J. Baker, "The design for the wall street journal-based csr corpus," *in Proc. ICSLP*, pp. 899–902, 1992.
[36] D. Hu, "100nonspeechenvironmentalsounds2004[online]," http://www. cse.ohio-state.edu/pnl/corpus/HuCorpus.html, 2004