Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

End-to-end Speech Intelligibility Prediction Using Time-Domain Fully Convolutional Neural Networks

Pedersen, Mathias; Kolbæk, Morten; Andersen, Asger Heidemann; Jensen, Søren Holdt; Jensen, Jesper

End-to-end Speech Intelligibility Prediction Using Time-Domain Fully Convolutional Neural Networks

Mathias B. Pedersen[1,*] , *Morten Kolbæk*[1,*], *Asger H. Andersen*[2], *Søren H. Jensen*[1], *Jesper Jensen*[1,2]

[1]Department of Electronic Systems, Aalborg University, Denmark
[2]Oticon A/S, Denmark

{mbp,mok,shj,jje}@es.aau.dk, {aand,jesj}@demant.com

Abstract

Data-driven speech intelligibility prediction has been slow to take off. Datasets of measured speech intelligibility are scarce, and so current models are relatively small and rely on hand-picked features. Classical predictors based on psychoacoustic models and heuristics are still the state-of-the-art. This work proposes a U-Net inspired fully convolutional neural network architecture, NSIP, trained and tested on ten datasets to predict intelligibility of time-domain speech. The architecture is compared to a frequency domain data-driven predictor and to the classical state-of-the-art predictors STOI, ESTOI, HASPI and SIIB. The performance of NSIP is found to be superior for datasets seen in the training phase. On unseen datasets NSIP reaches performance comparable to classical predictors.

Index Terms: speech intelligibility prediction, fully convolutional neural networks, deep learning

1. Introduction

Data-driven speech enhancement has garnered huge interest in the last decade with studies such as [1–6]. A more recent trend has been towards end-to-end solutions like [7–10], working fully in the time-domain. Most of these speech enhancement studies aim at enhancing speech intelligibility (SI), either in the evaluation or even as part of the objective. SI is a very relevant aspect of processed speech intended for human listeners, e.g. telecommunication systems and hearing assistive devices. Unfortunately, SI is time consuming to measure and hence speech intelligibility prediction (SIP) is of great importance to the field of speech enhancement in particular, and to the broader area of speech processing in general. SIP as a field however, has not seen the same rapid advancement in terms of data-driven methods as other fields in speech processing.

Presently, data driven SIP has only been attempted with relatively small datasets, and partially data-driven models using hand-engineered features [11–16]. Why is this? One of the main reasons is certainly that data-driven SIP is limited by data scarcity. In most other speech processing fields ground truth data is simply clean speech signals, which are relatively easily obtainable in bulk. Obtaining training data for SIP, however, requires time-consuming measurements of speech intelligibility through listening tests of individual noise/processing conditions. Thus the availability of speech data accompanied by subjectively measured SI is rather low.

Most state of the art SI-predictors like STOI [17], ESTOI [18], SIIB [19] and HASPI [20], are still not based on machine learning, but rather on psychoacoustic models and heuristics, and validated empirically using relatively small datasets with measured intelligibility. In spite of their non-data-driven design,

---
*Equal Contribution

these predictors have demonstrated excellent performance in a variety of noise and processing conditions, and remain among the most widely used. An overview of classical predictors is presented in [21]. It is, however, not fully understood exactly under which conditions these predictors perform well.

Some *data-driven* SI-predictors have been proposed, but they are all limited in one way or another. In [11–13] existing non-data-driven intelligibility predictors are used to either label the training data or as part of the architecture respectively. The systems in [14–16] are trained with measured intelligibility, though [14] uses data from a single listening test. These systems all rely on hand determined features, i.e. Mel frequency bands in [14], and 1/3-octave bands in [15, 16].

In this paper we propose and analyse the performance of an intrusive end-to-end speech deep neural network (DNN) intelligibility predictor. The network is a fully convolutional architecture inspired by U-Net [22] and resembles that used in a large body of literature including works involving speech enhancement (e.g. [7, 23, 24]). This network is trained and tested on speech and SI measurements of a wide variety of conditions from a range of listening tests. The network takes time-domain speech signals along with the corresponding clean speech as input and outputs SI-predictions as a function of time, and is thus an end-to-end data-driven SI-predictor. The architecture is explained in greater detail in Section 2 and the data and simulations are described in Section 3. The predictor is tested in a comparison with ESTOI, SIIB and HASPI, using the Pearson and Spearman correlation within each listening test. The results are presented in Section 4, and the conclusion in Section 5.

2. Data-driven Intelligibility Prediction

In this study we use a data-driven approach for speech intelligibility prediction. Specifically, we propose the neural speech intelligibility predictor (NSIP) model given by Fig. 1, which shows the architecture of an end-to-end intrusive speech intelligibility predictor based on fully convolutional neural networks.

2.1. Intrusive Speech Intelligibility Prediction

Intrusive SIP refers to the problem of estimating the SI of a noisy/processed speech signal, $x[t]$, using $x[t]$ itself and the corresponding clean speech signal, $s[t]$. Intrusive SI-predictors are classically more successful than their non-intrusive counterparts, which only rely on $x[t]$. Intrusive prediction can use $s[t]$ as a reference to measure how dissimilar $x[t]$ is to clean speech, while non-intrusive prediction requires a built-in model of generic clean speech in order to make such a comparison. This makes the classical intrusive predictors simpler and more robust. In transitioning to DNN's, the argument of simplicity
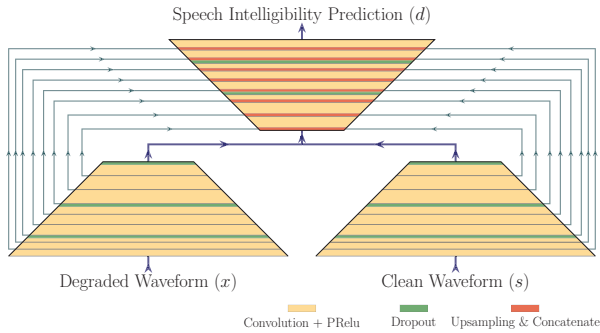
Figure 1: *Architecture of an intrusive neural speech intelligibility predictor based on fully convolutional neural networks. The predictor is trained end-to-end to estimate the sample-level speech intelligibility of a degraded speech waveform.*

Model	#filters in encoder layers $1-9$				#filters in decoder layers $10-18$				#Params (millions)
	$1-3$	$4-6$	$7-8$	9	$10-11$	$12-14$	$15-17$	18	
NSIP1	6	12	16	32	32	16	12	1	0.122M
NSIP2	8	16	24	64	64	24	16	1	0.349M
NSIP3	12	18	36	80	80	36	18	1	0.603M
NSIP4	12	24	48	96	96	48	24	1	0.946M
NSIP5	16	32	64	128	128	64	32	1	1.68M

Table 1: *Number of output filters in each layer of the NSIP-model given by Fig. 1 for five different configurations. All filters are 11 samples long.*

changes, because DNN's rely on their great parametric complexity in the first place. This makes non-intrusive architectures somewhat simpler, because they only need to work with one input rather than two. Intrusive architectures still have the potential to be more robust though, and because of the data scarcity, the extra clean speech input might be valuable.

The network architecture used in this paper is intrusive, since it receives the inputs, $s[t]$ and $x[t]$, which in this context are time-domain clean and noisy/distorted speech signals. The desired output is defined as a time domain piece-wise constant curve, $d[t]$, corresponding to measured SI of the input $x[t]$, as it is also done in [16]. The network output can then be integrated over time to produce an SI prediction for a particular span of time.

2.2. Neural Speech Intelligibility Prediction

The NSIP model depicted in Fig. 1 is based on a fully convolutional neural network architecture with 18 convolutional layers utilizing parameterized ReLU (PReLU) activation functions between the layers [25]. The model is inspired by U-Net [22] and follows an encoder-decoder methodology where skip-connections are applied between corresponding layers to allow data at various sample rates to flow between the encoder and decoder.

Differently from a standard U-net, the proposed model has two encoders, as shown in Fig. 1, one for the clean and one for the degraded speech waveforms, since intrusive speech intelligibility prediction can make use of both of these. Specifically, the two encoders each contain eight convolutional layers and the output of the two encoders, which contain compressed information about the clean and degraded speech signals, are concatenated and propagated to a joint decoder that performs the final SI prediction. The encoders both use a stride of two in each layer, except for the first layer where a stride of one is used. This drives the final dimension at the outputs of the encoders to be compressed with a factor of 256. Similarly, all layers in the decoder, except for the last layer, use upsampling with a factor of two, such that the final output has the same dimension as the inputs, which allows sample-level SI prediction.

To study how the number of parameters influence the SI performance of the proposed architecture, five NSIP models are trained and evaluated with a varying number of filters. The configurations of the individual NSIP systems are shown in Table 1. The number of parameters for the five models vary

from 0.122×10^6 to 1.68×10^6, which is comparable to the 0.224×10^6 parameters of a recently published frequency-domain technique [16] that will serve as an NSIP baseline in Sec. 4. Finally, all filters have a size of 11 samples.

The SIP-systems are trained to minimize the binary cross entropy between estimated and measured intelligibility using the ADAM optimizer [26] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and an initial learning rate of 0.0005, which is controlled by a learning rate schedule that reduces the learning rate with a factor of two, if the validation loss has not decreased for two epochs. Finally, during training, 20% dropout is applied for every third layer, and a batch size of 16 is used. Training is stopped, if the validation loss has not decreased for five epochs or a maximum of 200 epochs has elapsed.

The SIP-systems have been implemented using Keras[1] with a TensorFlow[2] backend and the python implementation of the trained NSIP-models, are available online[3], to allow interested readers, to use and evaluate the models further.

3. Experimental Design

To establish the potential of the proposed architecture in terms of predicting speech intelligibility of noisy/distorted speech, a series of experiments are conducted. In the following, the datasets used for training, validation, and test are presented.

3.1. Training, Validation and Test Data

Table 2 summarizes the ten datasets used for training, validating and testing the NSIP-models. The data consist of clean and noisy/distorted speech signals and measured SI scores, which are used as labels. Due to the number of datasets, space limitations make it impractical to give a detailed description of each listening test here. Since they are all well described in other works, we instead refer the interested reader to the respective sources. The datasets contain multiple talkers, languages, noise types and processing schemes. Classical predictors have shown varying performance on different subsets of these datasets, which is also verified in Section 4. There are significant differences in the size of these datasets, and Table 2 contains a breakdown of the size (files) and number of different acoustic conditions (cond.) in each dataset. Because of the limited amount of data, we do not attempt to balance the datasets by excluding data from the bigger datasets.

---

[1]https://keras.io/

[2]https://tensorflow.org/

[3]https://git.its.aau.dk/mok/neural_sip.git

Dataset		Training		Validation		Test	
No.	Ref.	#files	#cond.	#files	#cond.	#files	#cond.
0	[18]	564	60	60	58	60	58
1	[27]	6295	168	673	168	840	168
2	[17]	320	34	35	32	35	32
3	[15]	1744	327	77	76	318	299
4	[28]	784	24	96	24	96	24
5	[29]	439	18	54	18	54	18
6	[18]	3460	20	436	20	437	20
7	[30]	0	0	0	0	278	9
8	[31]	0	0	0	0	241	20
9	[32, 33]	0	0	0	0	64	52

Table 2: *Datasets used for training, validation and test. Each file corresponds to approx. 6.6s of speech. See references for further details regarding the general design of the datasets.*

3.2. Cross Validation

Datasets 0 – 6 have been split randomly into training, validation and test comprised of approximately 80, 10, and 10 % of the data, respectively. Each listening test condition has been split in this way, such that every condition is represented in the test set. Furthermore, due to the limited amount of test data available, 10-fold cross validation has been performed and for each split of the data into training, validation, and test, ten differently initialized sets of NN-weights have been trained. In other words, 100 models of each architecture have been trained. Finally, to demonstrate the performance in unseen conditions datasets 7 – 9 have been left out of the training and validation sets, and are used exclusively for testing. As such we distinguish between *seen* conditions, i.e. belonging to 0 – 6 and *unseen* conditions belonging to 7 – 9.

4. Experimental Results

4.1. End-to-end Data-driven Intelligibility Prediction

The NSIP-models defined in Table 1 have been evaluated using Spearman and Pearson correlation. The models were given the clean references and corresponding noisy/processed test data signals, and the predictions were integrated over each acoustic condition. Examples of these integrated predictions can be seen, compared to measured SI, in Figure 2. The Spearman and Pearson scores were then computed and are presented in Tables 3 and 4 with standard deviations from the cross-validation reported in parentheses. Spearman is a rank correlation and measures monotonicity between predictions and measurements, whereas Pearson correlation measures the linearity of their relationship. For each dataset the Spearman and Pearson correlation of the NSIP predictions are measured.

From Tables 3 and 4 it is seen that NSIP5 with 1.68×10^6 parameters reaches an average Spearman of .91 across seen conditions and .85 across unseen conditions, with corresponding average Pearson correlations of .91 across seen conditions and .85 across unseen conditions. The performance of NSIP5 is visualized for a few datasets in Figure 2.

4.2. Data-driven vs. Non-data-driven SIP

We compare the results from the NSIP-models on the test data with the classical predictors STOI, ESTOI, HASPI and SIIB, and a retrained network with the architecture of [16]. Simi-
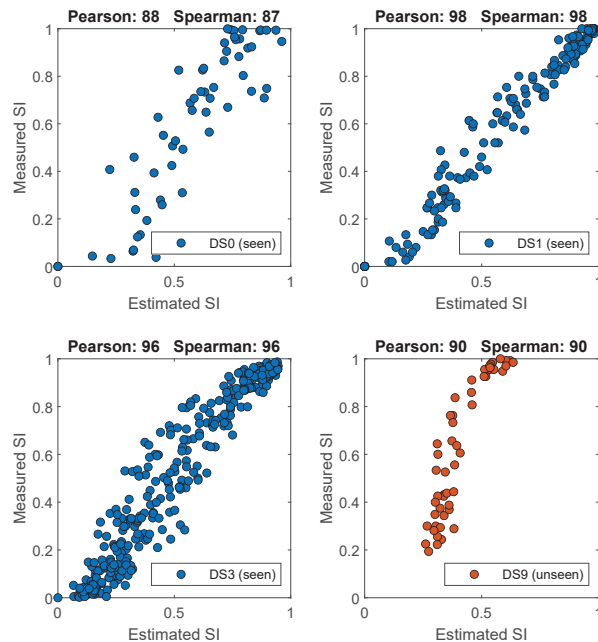

Figure 2: *Scatter plots showing relation between measured SI and estimated SI, estimated by the NSIP5 system, for seen datasets DS0, DS1 and DS3, as well as the unseen dataset DS9. The Pearson and Spearman correlations are scaled a factor of 100.*

lar to STOI and ESTOI, this architecture takes 1/3-octave band representations of s and x as inputs and outputs SI-predictions, and as such can be used as a frequency-domain benchmark. Tables 3 and 4 show the dataset-wise results in terms of Spearman and Pearson correlation respectively, for the NSIP-models and the classical predictors. We distinguish between the conditions which have and have not been seen by the NSIP-models during training, and report the average of the performance measures across these subsets as well. We stress that "seen" conditions are not training data, but distinct test data signals belonging to listening test conditions that also appear in the training set. In the case of Pearson correlation, a dataset dependent logistic curve is often fitted to the predictions before computing the correlation. This function has been used to map SI-predictions to measurements by [17, 19]. We do this for the classical predictors, and the Pearson correlations denoted by (fitted) in Table 4 thus measure the correlation in a logistic rather than linear sense. This increases their average Pearson correlation, but in the seen conditions, even with the added dataset-specific knowledge, they are still outperformed by the NSIP architectures, which has been given no such dataset-specific mapping.

The NSIP-models achieve better average performance, in terms of Spearman and Pearson correlation in seen conditions as compared to the classical predictors. Comparing the measures for the unseen datasets, NSIP is on par with the classical methods for datasets 7 and 9, but not dataset 8. Consequently, the average NSIP performance on the unseen datasets is lower than average performance of the classical predictors on the same datasets.

	Spearman ×100												
	Mean	Mean	Seen Data							Unseen Data			#Params
Predictor	(seen)	(unseen)	DS0	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9	(millions)
NSIP1 (time) :	82 (2.9)	85 (7.1)	76 (5.2)	96 (0.4)	78 (2.3)	93 (1.1)	57 (3.4)	74 (7.1)	98 (0.6)	97 (1.6)	77 (10.7)	80 (9.0)	0.122M
NSIP2 (time) :	85 (2.3)	82 (5.1)	84 (2.8)	97 (0.2)	81 (2.1)	95 (0.6)	64 (4.9)	76 (5.2)	98 (0.4)	98 (1.1)	64 (9.9)	85 (4.3)	0.349M
NSIP3 (time) :	88 (2.2)	83 (4.6)	87 (1.8)	98 (0.1)	82 (1.7)	96 (0.4)	73 (6.1)	80 (4.9)	99 (0.3)	97 (1.1)	64 (10.0)	87 (2.6)	0.603M
NSIP4 (time) :	89 (2.2)	85 (3.8)	87 (1.7)	98 (0.1)	83 (1.8)	96 (0.4)	81 (6.2)	81 (5.0)	99 (0.2)	98 (1.1)	69 (7.6)	87 (2.7)	0.946M
NSIP5 (time) :	91 (2.1)	85 (3.5)	88 (1.7)	98 (0.1)	84 (1.8)	96 (0.4)	87 (5.9)	83 (4.7)	99 (0.3)	97 (1.0)	70 (7.3)	89 (2.2)	1.68M
NSIP6 (freq):	88 (1.9)	74 (4.7)	79 (3.7)	97 (0.1)	81 (1.4)	96 (0.6)	82 (4.1)	83 (3.0)	97 (0.4)	96 (1.9)	70 (5.1)	56 (7.2)	0.224M
STOI:	74	93	47	96	60	81	57	83	98	95	96	87	–
ESTOI:	78	92	82	96	49	84	56	86	96	98	95	85	–
HASPI:	71	88	62	78	50	93	64	65	84	98	96	70	–
SIIB:	80	96	73	91	39	93	75	94	98	98	97	94	–

Table 3: *Spearman correlation for NSIP models and classical non-data-driven SIP techniques. NSIP1-5 are time-domain models configured according to Fig. 1 and Table 1 and NSIP6 are an frequency-domain baseline model from [16]. All models are trained with data according to Table 2. The score are mean scores computed based on 10-fold cross validation and the scores in parenthesis are standard deviations.*

	Pearson Correlation ×100												
	Mean	Mean	Seen Data							Unseen Data			#Params
Predictor	(seen)	(unseen)	DS0	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9	(millions)
NSIP1 (time) :	84 (2.7)	83 (7.0)	75 (4.5)	96 (0.4)	77 (2.5)	93 (1.1)	77 (3.7)	76 (5.9)	97 (0.6)	95 (2.1)	76 (10.3)	77 (8.7)	0.122M
NSIP2 (time) :	88 (1.7)	80 (6.1)	83 (2.8)	97 (0.3)	80 (1.7)	95 (0.6)	87 (1.6)	79 (4.7)	98 (0.4)	97 (1.1)	62 (13.0)	83 (4.2)	0.349M
NSIP3 (time) :	90 (1.4)	81 (5.7)	86 (2.1)	98 (0.2)	81 (1.4)	96 (0.4)	89 (1.0)	82 (4.2)	98 (0.2)	96 (1.4)	62 (12.8)	85 (2.8)	0.603M
NSIP4 (time) :	91 (1.2)	84 (4.1)	87 (1.9)	98 (0.2)	82 (1.4)	96 (0.4)	90 (0.8)	83 (3.7)	99 (0.2)	97 (1.3)	69 (8.5)	86 (2.7)	0.946M
NSIP5 (time) :	91 (1.1)	85 (3.7)	89 (1.6)	98 (0.1)	83 (1.2)	96 (0.4)	91 (0.8)	85 (3.5)	99 (0.2)	96 (1.3)	71 (8.0)	87 (1.7)	1.68M
NSIP6 (freq):	89 (1.2)	73 (5.2)	77 (3.8)	97 (0.1)	79 (1.1)	96 (0.6)	91 (0.7)	86 (2.1)	98 (0.2)	93 (2.0)	70 (7.1)	57 (6.5)	0.224M
STOI:	77	92	51	91	56	78	80	85	98	98	89	90	–
ESTOI:	79	92	77	93	44	80	81	86	95	97	93	86	–
HASPI:	62	80	42	77	45	85	37	69	81	91	74	76	–
SIIB:	77	88	62	85	32	80	89	95	94	96	77	90	–
STOI (fitted):	78	96	51	96	58	80	76	85	99	99	96	91	–
ESTOI (fitted):	81	94	83	95	45	82	78	87	97	100	95	88	–
HASPI (fitted):	65	89	61	77	45	88	36	70	80	97	93	78	–
SIIB (fitted):	82	97	74	90	33	92	92	95	98	99	95	96	–

Table 4: *As Table 3 but for Pearson correlation.*

4.3. Frequency-domain Data-driven SIP

In order to judge the potential advantage of an end-to-end architecture, we compare NSIP to the architecture of [16], which takes 1/3-octave band transformed speech signals as inputs, similar to STOI and ESTOI. This architecture has been retrained on the same data as the proposed time-domain NSIP architecture. This is done to gauge the advantage of NSIP's access to the full information in the time-domain. As was the case for the time-domain architecture, the frequency-domain architecture is trained and tested on the ten cross validation data-splits. The test results are shown in the rows labelled NSIP6 (freq) in Tables 3 and 4. It appears that the time-domain architectures of similar parameter size perform slightly better on average in terms of Spearman and Pearson on the unseen Datasets 7 and 8, and significantly better on Dataset 9. This could be due to the loss of information in the 1/3-octave band transform employed in NSIP6. On the seen datasets the frequency-domain architecture performs as well as NSIP3 and 4.

5. Conclusion

We proposed a time-domain neural speech intelligibility predictor (NSIP) based on a fully convolutional neural network architecture, for intrusive speech intelligibility prediction. This network was trained on seven listening test datasets and tested on ten. Performance was evaluated in terms of Spearman and Pearson correlation, and compared to the classical predictors STOI, ESTOI, HASPI and SIIB, and a retrained frequency-domain architecture, [16]. The NSIP architectures showed the best performance on the seven seen datasets, but were outperformed by the classical predictors on one of the unseen datasets. The frequency-domain architecture was found to reach performance similar to that of larger, in terms of parameters, time-domain architectures, with much fewer parameters.

6. References

[1] K. Han and D. Wang, "A classification based approach to speech segregation," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3475–3483, 2012.

[2] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[3] E. W. Healy, S. E. Yoho, J. Chen, Y. Wang, and D. Wang, "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," *The Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. 1660–1669, 2015.

[4] M. Kolbæk, Z. H. Tan, and J. Jensen, "Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, 2017.

[5] M. Kolbæk, Z. Tan, and J. Jensen, "On the Relationship Between Short-Time Objective Intelligibility and Short-Time Spectral-Amplitude Mean-Square Error for Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 283–295, 2019.

[6] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[7] A. Pandey and D. Wang, "A New Framework for CNN-Based Speech Enhancement in the Time Domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

[8] S. W. Fu, T. W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-End Waveform Utterance Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 570 – 1584, 2018.

[9] S. R. Park and J. Lee, "A Fully Convolutional Neural Network for Speech Enhancement," in *Proc. Interspeech*, 2017, pp. 1993–1997.

[10] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On Loss Functions for Supervised Monaural Time-Domain Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 1, pp. 825–838, 2020.

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

[12] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-HMM-based non-intrusive speech intelligibility prediction," *ICASSP*, pp. 624–628, Mar. 2016.

[13] P. Seetharaman, G. J. Mysore, P. Smaragdis, and B. Pardo, "Blind Estimation of the Speech Transmission Index for Speech Quality Prediction," *ICASSP*, pp. 591–595, Apr. 2018.

[14] K. Kondo, K. Taira, and Y. Kobayashi, "Binaural speech intelligibility estimation using deep neural networks," *Interspeech*, pp. 1858–1862, Sep. 2018.

[15] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1925–1939, 2018.

[16] M. B. Pedersen, A. H. Andersen, S. H. Jensen, and J. Jensen, "A Neural Network for Monaural Intrusive Speech Intelligbility Prediction," *ICASSP*, pp. 336–340, May 2020.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[18] 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.

[19] S. V. Kuyk, W. B. Kleijn, and R. C. Hendriks, "An Instrumental Intelligibility Metric Based on Information Theory," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 115–119, Jan. 2018.

[20] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Perception Index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, 2014.

[21] T. H. Falk, V. Parsa, J. 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: Advantages and limitations of existing tools," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, 2015.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. MICCAI*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., 2015, pp. 234–241.

[23] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Proc. INTERSPEECH*, 2017, pp. 3642–3646.

[24] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On Loss Functions for Supervised Monaural Time-Domain Speech Enhancement," 2019. [Online]. Available: http://arxiv.org/abs/1909.01019

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *Proc. ICCV*, 2015, pp. 1026–1034.

[26] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. ICLR (arXiv:1412.6980)*, 2015.

[27] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1415–1426, 2009.

[28] T. Bentsen, A. A. Kressner, T. Dau, and T. May, "The impact of exploiting spectro-temporal context in computational speech segregation," *J. Acoust. Soc. Am.*, vol. 143, no. 1, pp. 248–259, Jan. 2018.

[29] C. H. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 225–228, Mar. 2013.

[30] A. H. Moore, J. M. de Haan, M. S. Pedersen, P. A. Naylor, M. Brookes, and J. Jensen, "Personalized signal-independent beamforming for binaural hearing aids," *J. Acoust. Soc. Am.*, vol. 145, May 2019.

[31] A. H. Andersen, J. M. d. Haan, Z. H. Tan, and J. Jensen, "Predicting the Intelligibility of Noisy and Nonlinearly Processed Binaural Speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 1908–1920, 2016.

[32] W. B. Kleijn and R. C. Hendriks, "A simple model of speech communication and its application to intelligibility enhancement," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 303–307, Mar. 2014.

[33] R. C. Hendriks, J. B. Crespo, J. Jensen, and C. H. Taal, "Optimal near-end speech intelligibility improvement incorporating additive noise and late reverberation under an approximation of the short-time sii," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 5, pp. 851–862, May 2015.