BLIND ESTIMATION OF THE SPEECH TRANSMISSION INDEX FOR SPEECH QUALITY PREDICTION

Prem Seetharaman[⋆†], *Gautham J. Mysore*[†], *Paris Smaragdis*[‡†], *Bryan Pardo*[⋆]

⋆ Northwestern University - Evanston, IL
† Adobe Research - San Francisco, CA
‡ University of Illinois at Urbana-Champaign - Champaign, IL

ABSTRACT

The speech transmission index (STI) of a listening position within a given room indicates the quality and intelligibility of speech uttered in that room. The measure is very reliable for predicting speech intelligibility in many room conditions but requires an STI measurement of the impulse response for the room. We present a method for blindly estimating the STI without measuring or modeling the impulse response of the room using deep convolutional neural networks. Our model is trained entirely using simulated room impulse responses combined with clean speech examples from the DAPS dataset [1] and works directly on PCM audio. Our experiments show that our method predicts true STI with a high degree of accuracy – an average error of under 4%. It can also distinguish between different STI conditions to a level of granularity that is comparable to humans.

Index Terms— Speech quality, speech enhancement, speech transmission index

1. INTRODUCTION

The study of speech intelligibility is the study of how comprehensible speech is to listeners, given environmental conditions. These conditions include background noise level, reverberation characteristics (e.g. reverberation time), and distortions in the sound producing equipment (e.g. low quality loudspeaker). Many measures have been proposed for objective evaluation of speech intelligibility, such as PESQ [2], PEAQ [3], and STOI [4]. One of the most successful measures to date is the speech transmission index (STI) [5]. The speech transmission index of a listening position within a given room very reliably indicates the quality and intelligibility of speech uttered in that room [6].

The idea behind the speech transmission index is that the effect an environment has on the spectro-temporal modulations of speech is correlated with speech intelligibility. If these modulations are kept intact, the room has a high speech transmission index. If the modulations are destroyed or smeared, the speech transmission index is low. Modulations of speech can be destroyed by reverberation or excessive background noise.

The speech transmission index ranges from 0 (worst) to 1 (best). This range covers a wide variety of acoustic conditions from large public spaces like sports stadiums (around .3 to .6) to bedrooms and offices (around .8 to .9) all the way up to professional recording studios (around .97 and above). The measure is very reliable for predicting speech intelligibility in many room conditions. In informal listening tests, we found that STI can be used to distinguish pleasant recording scenarios (such as those on professional radio programs) from amateur recordings (such as podcasts recorded in a living room).

The speech transmission index is measured by estimating the transfer function of a given room with respect to given speaker and listener positions [7]. This is a laborious manual process that can by done by creating a signal that mimics the modulations of speech in different frequency bands, playing it through a high quality loudspeaker, and recording the output with a high quality microphone. This process takes up to 15 minutes in good conditions [8]. It can alternatively be computed from a measurement of the room impulse response, whose measurement is also laborious [9]. Further, It is not always possible to take an STI measurement of a space (e.g. in public spaces like a subway platform). Therefore, the STI for most prerecorded audio cannot be calculated.

In this work, we present a method to reliably estimate the speech transmission index from speech recordings, circumventing the need to take an STI measurement with specialized sound sources (modulated noise) and equipment (high-quality microphones and loudspeakers). To do this, we train a convolutional neural network that computes a regression from time series audio of speech to the speech transmission index for that room. Applications of our system include identifying high-quality speech data in large unlabeled speech datasets (e.g. LibriVox recordings), informing users of recording software of problems in their recording setup, or diagnosing problems for speech recognition systems (e.g. telling users to

This work was done while on an Adobe Research internship. Paris Smaragdis was supported through NSF grant #1451380.

move their smart home device to locations where the speech transmission index is higher for more reliable usage).

2. RELATED WORK

The most closely related work to our own is a method for blind estimation of speech transmission index by Unoki et al. [10]. The speech transmission index is easily calculated given the impulse response for a given room. The crux of their method is to estimate the speech transmission index of a room by computing it from an approximation of the impulse response of the room. The approximation is derived using a generalization of Schroeder's room impulse response model ([11]) and has three parameters: the reverberation time, the gain factor, and the order of the impulse response. Estimating these three parameters is constrained by the behavior of the spectrotemporal modulations of the observed, reverberant speech. Their method relies on accurate estimation of these three parameters and a realistic model for room impulse responses.

In contrast, our method makes no assumptions about the model of the room impulse response. Their model is constrained to the accuracy of generalized Schroeder's room impulse response. Our method only assumes that the observed signal is that of speech. We use a deep neural network that regresses from the reverberant speech to the speech transmission index. Additionally, their system is only tested and built for acoustic conditions with STIs between .4 and .8 whereas our system leverages a broader spectrum of STIs all the way up to .99 (professional recording studios). This broader spectrum includes STIs corresponding to excellent recordings (e.g. recordings from professional radio programs) and amateur recordings (e.g. recordings from amateur podcast producers).

The method in [10] and our proposed method are the only existing methods for blind estimation of the speech transmission index we are aware of. A closely related measure is the speech-to-reverberation modulation energy ratio [12], which leverages similar assumptions to the method in [10]. However, many methods have been developed for blind estimation of other room parameters, such as reverberation time ([13], [14]). Xiao et al. [15] use a deep neural network that estimates reverberation time from spectrogram patches. Our work instead estimates the speech transmission index, which has a more reliable relationship to speech quality [7] (see Figure 1). In [16], the authors implement a system that estimates the source-to-distortion ratio (SDR) using deep neural network regression. The technique we employ here is similar but estimates the speech transmission index rather than SDR and uses a simpler network with fewer parameters.

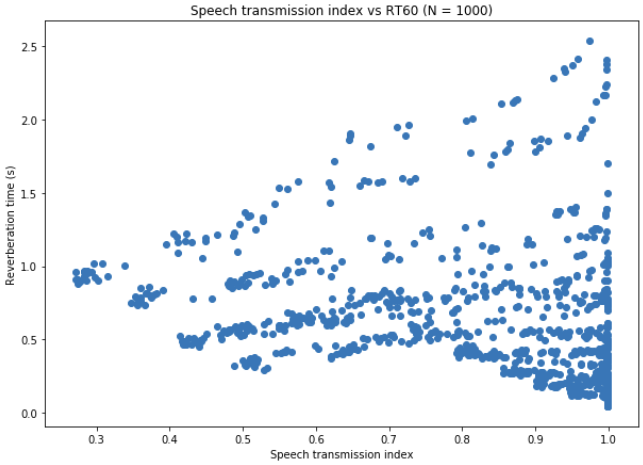

Fig. 1. Speech transmission index versus RT60, a common measurement of reverberation. The graph is generated from our synthetic impulse response dataset (Section 3.1). The two measures are poorly correlated. Long RT60 can still have high STIs and short RT60 can have low STI.

3. METHOD

3.1. Training data

Our training data is based on speech recordings from the DAPS (device and produced speech) dataset [1]. The clean version of the recordings in the DAPS dataset consists of twenty speakers (ten male, ten female) reading five excerpts from public domain stories (about 14 minutes per speaker - 280 minutes for the entire dataset). We took the clean recordings from DAPS and split them randomly into training and testing sets, each consisting of 10 speakers (5 male and 5 female - 140 minutes of clean speech). These recordings were segmented into 1 second chunks with no overlap. Any 1-second chunks that don't contain any speech were removed. The recordings from DAPS were downsampled to 16000 Hz to reduce computational cost.

We performed data augmentation to increase the amount of data to train our model. We created a library of 1000 artificial impulse responses using a room impulse simulator [17]. These impulse responses were generated across a variety of room conditions. The room dimensions varied from 5 meters to 20 meters along each axis (height, width, and depth). The absorption coefficients for each wall was chosen from the set $[.01, .1, .3, .5]$. The room impulse responses were generated using the image-source method [18]. The source (speech) was placed at 1/3 the height, width, and depth of the room. Virtual microphone locations were sampled at varying distances from the source. Impulse responses were computed for every microphone-source pair in every room. This procedure resulted in 1000 artificial impulse responses. 500 of these were placed in a training set and the other 500 were placed in a

testing set.

For each impulse response, we computed the speech transmission index using the method described in [5]. We then used the clean speech from DAPS and the generated impulse responses with corresponding speech transmission indices to create a dataset. We generated this dataset on the fly during training as follows. First, a random selection of n 1-second audio excerpts were selected from DAPS. Then, a random selection of n impulse responses were selected from the impulse response dataset. Each 1-second audio chunk of an excerpt was convolved with the corresponding impulse response to produce a reverberant speech signal. The reverberant speech signal was then paired with the speech transmission index corresponding to the impulse response used to generate the reverberant speech, forming a labeled example (audio signal and speech transmission index).

3.2. Network architecture

We used a fully convolutional neural network with 40095 trainable parameters. Our network takes a one-second clip of the PCM audio (16,000 samples) as input and outputs an estimate of the STI. The network architecture is shown in Table 1. The first convolutional layer in the network computes a spectrogram representation of the audio with 128 filters of length 128 samples (8ms at 16kHz) with a hop size of 64 samples. The weights of this layer are initialized with a Fourier basis (sine waves at different frequencies) and are updated during training to find an optimal spectrogram-like transform of the data for the task. After this learned time-frequency representation is obtained, it is passed through a series of 2D convolutions, leaky ReLU units, and batch normalization layers. The size of the input is cut in half at each layer until 1 second of audio data maps onto a single number. The output of the last convolutional layer is passed through a sigmoid activation unit to map the output between 0 and 1 (lower and upper bound for STI, respectively).

3.3. Training the model

The network was trained using the ADAM optimizer [20] with a loss function of mean squared error between the predicted and ground truth speech transmission index. We used a learning rate of .001 and the model was trained for 200 epochs with a batch size of 32. An epoch was a pass over every clean speech sample in our training dataset, convolved with some set of impulse responses from the generated impulse responses. In total, this makes for $2,703,168$ possible training examples (with each example being 1 second of reverberant speech) in our data generation approach, which randomly chooses speech data and impulse response data to train with. For 200 epochs, this corresponds to roughly 322 hours of training data. We implemented our models in PyTorch.

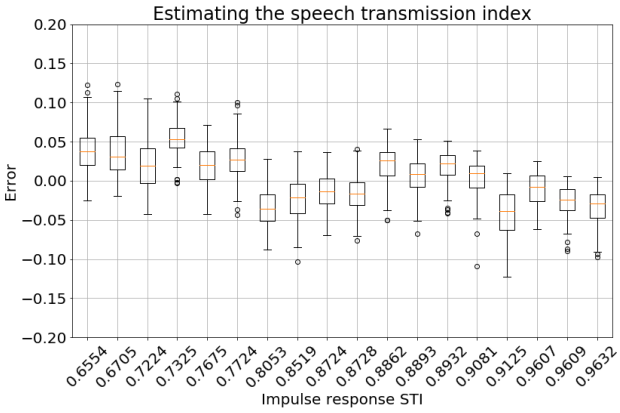

Fig. 2. Error in predicting the speech transmission index for each impulse response in the dataset from the REVERB challenge [19]. A value of 0 indicated perfect performance. The network tends to overestimate lower STI values and underestimate higher STI values. The overall root mean square error was 0.037.

4. EVALUATION

We evaluated our model by using the test set of speakers from DAPS (see Section 3.1) and the set of 18 real-world impulse responses from the REVERB challenge [19]. We use these real-world impulse responses to show that our model did not simply memorize the conditions of the room acoustics simulator used for training the model. The synthetic impulse responses used for training were perfect impulse responses, whereas the ones used for testing were collected in rooms using imperfect impulses (e.g. balloon pops). Instead of using 1 second audio excerpts from the speech dataset as in training, we use 5 second audio excerpts. We do this to show our system working on longer more realistic audio excerpts. Each 5 second excerpt is convolved with one of the 18 impulse responses from the REVERB challenge. These 5 second excerpts are passed to the model which outputs estimates of the speech transmission index over the course of the excerpt. To get a single speech transmission index for the entire excerpt, we take the mean of all of the estimated speech transmission indices for the entire recording. We construct 2000 testing examples.

We then measured how close the STI estimate using our method is to the ground truth STI. Figure 2 shows the prediction performance as a function of the speech transmission index (STI). It shows a tendency to overestimate lower STI conditions and underestimate higher STI conditions. Our method estimates the speech transmission index within 3.7% of the actual speech transmission index on average. A competing method from Unoki et al. [10] reports a similar experiment (with different speech data and impulse responses). Due to unavailable code and unavailable speech

Layer type	Input	Conv (1D)	Conv (1D)	Conv2D	Conv2D	Conv2D	Conv2D	Conv2D
# of Filters	-	128	128	8	16	32	1	1
Output Shape	$(N, 1, 16000)$	$(N, 128, 253)$	$(N, 128, 253)$	$(N, 8, 253)$	$(N, 16, 111)$	$(N, 32, 40)$	$(N, 1, 5)$	$(N, 1)$
Filter Size/Stride	-	128, 64	5, 1	(128, 1), (128, 1)	(1, 32), (1, 2)	(1, 32), (1, 2)	(1, 32), (1, 2)	(1, 5)
Activation Function	-	-	-	LeakyReLU	LeakyReLU	LeakyReLU	-	Sigmoid
Notes	1 sec. audio	Fourier init.	Spectrogram smoothing	Batch norm before LeakyReLU	Batch norm before LeakyReLU	Batch norm before LeakyReLU	-	-

Table 1. Network architecture. The input to the network is 1 second of PCM audio of the reverberant speech. It is passed through a series of convolutional layers. The first convolutional layer generates a spectrogram-like representation of the audio. A series of 2D convolutional layers is then applied to the representation. Each layer halves the size of the representation until just 1 number is output for every second of audio. The network has 40095 trainable parameters. N is the batch size.

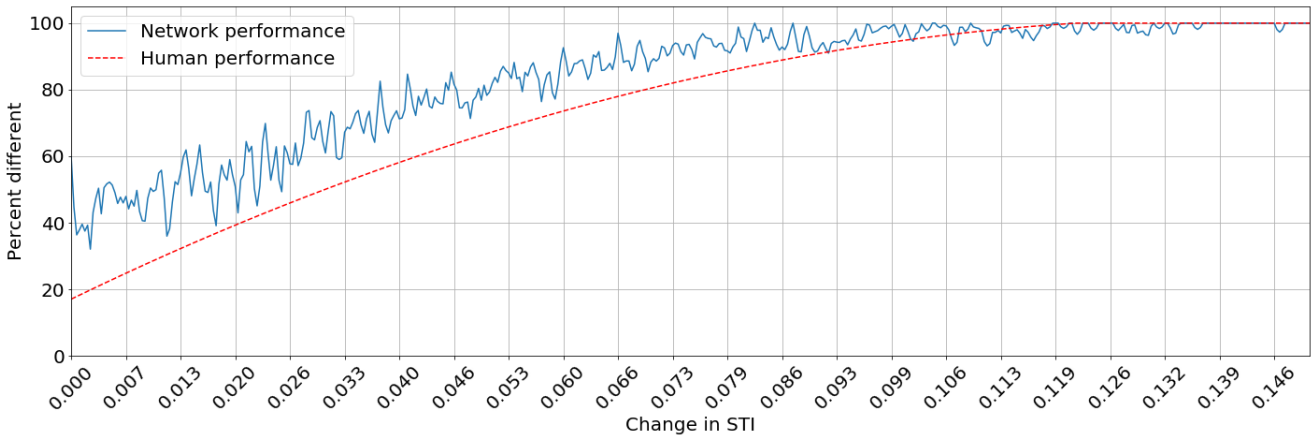

Fig. 3. Network performance in distinguishing STI conditions versus human performance (taken from the polynomial regression published in [6]). The just noticeable difference (50%) in STI is around the same for the network (.026) and for humans (.03). The recommended increase in speech transmission index to have an obvious impact on speech quality is around .1 for both.

and impulse response data, we were unable to directly compare our method with their system (or a working reproduction of their method). Their experiment reports an RMS error of .049 on reverberant speech signals. While the experiments use different data, the reported root mean squared errors on similar tasks suggests our method has a 24% improvement over a competing method. Additionally, the root mean square error of our method (.037) is close to the reported just noticeable difference (.03) for the speech transmission index [6].

We then tested whether our system performs well when distinguishing speech transmission index conditions from each other. Figure 3 shows the network performance in a just noticeable difference (JND) experiment on the speech transmission index, overlaid with human performance on a similar experiment (taken from [6]). In this experiment, the network is queried with pairs of reverberant speech examples that have different corresponding STIs. We use the estimated STIs output by the network to decide which of the reverberant speech examples has the higher STI. Figure 3 shows that as the difference gets larger, the network performs better.

In [6], the authors report the results of a similar experiment done with humans, as well as a regression for that experiment. The experiment was reported using clarity (C50) but they show a way to convert between C50 and STI. They found that humans have a just noticeable difference of .03 for the speech transmission index and get around 100% accuracy at a difference of .1. We see similar behavior for our model, with a just noticeable difference of .026 and perfect accuracy around a difference of .1.

5. CONCLUSION

The speech transmission index is a salient measure of how intelligible speech is in a given room. Traditionally, measuring the STI is a laborious task that requires access to the room. In this work, we have presented a method for blindly estimating the speech transmission index using convolutional neural networks. The method is trained entirely using simulated room impulse responses combined with clean speech examples from the DAPS dataset [1]. In experiments, the network performance is on par with human performance on a similar task. Perhaps models trained to predict other perceptual metrics such as PESQ [2], PEAQ [3], STOI [4] from PCM audio could be used to estimate just noticeable differences for humans. Our method also performs well in absolute terms, estimating the speech transmission index within 3.7% of the actual speech transmission index on average.

6. REFERENCES

[1] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?a dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1006–1010, 2015.

[2] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP'01)*, vol. 2, pp. 749–752, IEEE, 2001.

[3] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "Peaq-the itu standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.

[4] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4214–4217, IEEE, 2010.

[5] T. Houtgast and H. J. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Acta Acustica United With Acustica*, vol. 28, no. 1, pp. 66–73, 1973.

[6] J. Bradley, R. Reich, and S. Norcross, "A just noticeable difference in c 50 for speech," *Applied Acoustics*, vol. 58, no. 2, pp. 99–108, 1999.

[7] T. Houtgast, H. Steeneken, W. Ahnert, L. Braida, R. Drullman, J. Festen, K. Jacob, P. Mapp, S. McManus, K. Payton, *et al.*, "Past, present and future of the speech transmission index," *Soesterberg: TNO*, p. 73, 2002.

[8] 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.

[9] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 249–262, 2002.

[10] M. Unoki, K. Sasaki, R. Miyauchi, M. Akagi, and N. S. Kim, "Blind method of estimating speech transmission index from reverberant speech signals," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, pp. 1–5, IEEE, 2013.

[11] M. R. Schroeder, "Integrated-impulse method measuring sound decay without using impulses," *The Journal of the Acoustical Society of America*, vol. 66, no. 2, pp. 497–500, 1979.

[12] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[13] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. OBrien Jr, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[14] J. Y. Wen, E. A. Habets, and P. A. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 329–332, IEEE, 2008.

[15] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "Learning to estimate reverberation time in noisy and reverberant rooms," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[16] E. Manilow, P. Seetharaman, F. Pishdadian, and B. Pardo, "Predicting algorithm efficacy for adaptive multi-cue source separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2017. WASPAA'17.*, 2017.

[17] H. Pan, R. Scheibler, E. Bezzam, I. Dokmanić, and M. Vetterli, "Frida: Fri-based doa estimation for arbitrary array layouts," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 3186–3190, IEEE, 2017.

[18] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[19] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[20] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.