Non-Intrusive Signal Analysis for Room Adaptation of ASR Models

Ge Li[1], Dushyant Sharma[2], and Patrick A. Naylor[3]

[1]Nuance Communications, Canada
[2]Nuance Communications Inc., USA
[3]Imperial College London, UK

***Abstract*—We present a new deep-learning-based non-intrusive signal assessment method (NISA+) that performs a joint estimation of a large set of speech signal parameters, including those related to reverberation (C_{50}, DRR, reflection coefficient and room volume), background noise (SNR), perceptual speech quality (PESQ), speech intelligibility (ESTOI), voice activity detection, and speech coding (codec presence and bitrate). We show that neural embedding based combination of spectral features with an LSTM and modulation features with a convolution neural network enable NISA+ to achieve state of the art performance. Particularly, for non-intrusive PESQ and C_{50} estimation, we show around 15% relative reduction in estimation error compared to our previous best results . We also show that NISA+ can be used to perform targeted data augmentation for generating training data for ASR that matches the signal characteristics extracted from a small sample of data recorded in a target room acoustic environment. We show that a 9.6% word error rate reduction can be achieved relative to an ASR model trained with random augmentation.**

I. Introduction

In real world applications, clean speech can be corrupted by many factors including room reverberation, additive noise and coding artifacts, degrading the quality and intelligibility of the signal. The estimation of parameters characterizing these corrupting factors, as well as the perceived quality and intelligibility of the speech, has important applications including Automatic Speech Recognition (ASR) [1], [2], audio forensics [3], text-to-speech (TTS) [4] and speaker diarization [5]. A number of intrusive methods for estimating such parameters have been proposed. However, in real world deployments, the requirement of a reference signal in such non-intrusive methods is typically not fulfilled. This has led to the development of data-driven, non-intrusive methods in the recent decade.

The process of room reverberation can be modeled as a convolution between anechoic speech and a Room Impulse Response (RIR) [6]. The effects of reverberation have typically been characterized by the following intrusive parameters (extracted from an RIR): reverberation time (T_{60}), clarity index (C_{50}) and Direct-to-reverberant ratio (DRR) [6]. In addition, a number of parameters can be defined for the simulation of RIRs, including room volume and reflection coefficients for reflective surfaces in a room. In [7], Parada et al. propose a non-intrusive method to estimate T_{60} and DRR based on the Bidirectional Long Short Term Memory (BLSTM) architecture and show good performance when tested on the Acoustic Characterization of Environments (ACE) challenge [8] data. Furthermore, in [9], [10], the authors showed that the C_{50} metric was most highly correlated with ASR performance, a claim further validated in [11], where a method for joint estimation of C_{50}, Signal to Noise Ratio (SNR), PESQ and Voice Activity Detection (VAD) was proposed. A non intrusive reverberation time estimator that uses a Convolution Neural Network (CNN) architecture was presented [12].

Typically, speech is encoded via a codec to reduce the transmission bandwidth. In [13], Scholz et al. present a codec identification method based on Spectral Harmonic Decomposition (SHD) that achieves a classification accuracy of over 92%. A more recent study [14] presents a non-intrusive algorithm for codec bit-rate detection that achieves an accuracy of 95.4% compared to 76.4% with the baseline algorithms.

The non-intrusive estimation of perceived speech quality is a challenging task due to its subjective nature. Target parameters used in previous speech quality assessment literature include the Mean Opinion Score (MOS), the intrusive PESQ score [15] and POLQA score [16]. In [17], the authors propose a deep-learning-based non-intrusive method to estimate MOS directly. They first train an auto-encoder for reconstructing the input signal spectrum and then use the latent variables learnt by the auto-encoder as input features to a multilayer perceptron (MLP) that predicts MOS. More recently, Sharma et. al. [18] proposed an LSTM-based method that estimates the intrusive POLQA score, combining Mel-Frequency Cepstal Coeficients (MFCC) and a compressed modulation domain feature set. They show that their proposed method can predict POLQA with a Mean Absolute Error (MAE) of 0.21. An early non-intrusive estimator for speech intelligibility was proposed in [19] but this remains a challenging task.

In the literature, non-intrusive methods typically estimate either speech quality parameters, room acoustics or codec parameters individually. Our recent work [11] proposed a speech assessment method–NISA (Non-Intrusive Speech Analysis), which jointly estimates C_{50}, SNR, VAD and PESQ using Mel Filterbank features and a CNN architecture.

In this paper we propose a novel configuration of Mel filterbank and modulation spectrum features using their neural

Ge was an intern at Nuance during the course of this work.

representations as obtained with recurrent and convolution network architectures, respectively. We also extend the NISA method to perform a wider range of speech assessment tasks, including codec detection (i.e. whether the input speech is coded), bit rate estimation, speech intelligibility, room volume and reflection coefficient estimation. Finally, we use the new NISA+ method for guiding the data augmentation process for training of an ASR system and show how it can be used to perform room adaptation and achieve a nearly 10% word error rate reduction when compared with random augmentation.

II. NISA Method

Here we describe the NISA+ method, starting with the spectral and modulation features, followed by neural network architectures, strategies for the fusion of the two feature embeddings and the training setup.

A. Speech Features

1) Mel Filterbank Coefficients: The Mel Filterbank Coefficients (MFB) are a popular spectral feature set that is used in many speech signal processing applications and are motivated by human auditory perception of speech [20]. These features are derived by applying multiple triangular filters on a Mel-Scale to the power spectrum calculated from the Short-Time-Fourier-Transform (STFT). We use a frame size of 20 ms with a 5 ms time increment and apply 80 Mel filters to the power spectrum.

2) Modulation domain features: The second feature set is based on the Modulation Spectrum (MS), which captures the modulation information in speech. It has been shown that linguistic information is primarily carried in the low-frequency modulations of speech and MS features have been successfully used in many applications including speech recognition [21] and modeling speech intelligibility [22]. The MS features are obtained by applying two successive STFTs to the speech signal as described in [18]. We use the same acoustic frame size and increment as the MFB feature extraction, with a modulation frame size of 400 ms and modulation step size of 200 ms. Given the selected acoustic step-size, the sampling frequency of the modulation signal is 200 Hz.

B. Neural Network Architectures

The first architecture used in this study is the Long Short Term Memory (LSTM) [23] network, which is a Recurrent Neural Network (RNN) structure designed to capture temporal dependencies in sequential data. Our LSTM structure is composed of an input layer followed by three hidden layers, arranged in a 108×54×27 cell topology, for each time-step. The second architecture we explore is based on a CNN structure proposed in [24] for speech presence detection that is inspired by WaveNet [25]. In our implementation, we use 8 layers of causal gated 1D convolution with [16,8,8,16,16,16,16,32] filters. This is followed by a dropout layer followed by a flattening operation. These architectures were determined experimentally and the best performing system uses MS features as the input to this CNN system and MFB features to the LSTM. The output from these two are fused and then input to a dense layer that is connected to each of the estimation task workers, as described in more detail in the following and depicted in Fig. 1.

1) Feature Fusion: We investigate two strategies for combining the information from MFB and MS features. The features are processed by an intermediate neural network, the outputs of which are fused in different ways. The first fusion strategy is a direct concatenation (Fig. 1a), where the representations or embeddings learnt from MFB and MS are concatenated directly, as follows:

$$\mathbf{X}_{Fused1} = [\mathbf{X}_{MFB}; \mathbf{X}_{MS}],$$

where $\mathbf{X}_{MFB} \in \mathbb{R}^{d_1}$ and $\mathbf{X}_{MS} \in \mathbb{R}^{d_2}$ are the embedding vectors extracted by a certain neural-network from MFB and MS respectively and $\mathbf{X}_{Fused1} \in \mathbb{R}^{d_1+d_2}$ is the fused embedding that combines the information from both features. The second fusion strategy (Fig. 1b) implements a gating block before concatenating the two feature embeddings:

$$\begin{aligned}
\mathbf{X}_{Fused1} &= [\mathbf{X}_{MFB}; \mathbf{X}_{MS}] \\
w_1 &= \sigma(\mathbf{W}_1^T \mathbf{X}_{Fused1} + b_1) \\
w_2 &= \sigma(\mathbf{W}_2^T \mathbf{X}_{Fused1} + b_2) \\
\mathbf{X}_{Fused2} &= [w_1 * \mathbf{X}_{MFB}; w_2 * \mathbf{X}_{MS}]
\end{aligned}$$

where $\mathbf{X}_{Fused1} \in \mathbb{R}^{d_1+d_2}$ is an intermediate vector which will be used to calculate the weighting. $\mathbf{W}_1 \in \mathbb{R}^{d_1 \times 1}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_2 \times 1}$ are learned parameters. In this strategy, the model will learn to control the relative weighting between the representations learnt from two distinct speech features.

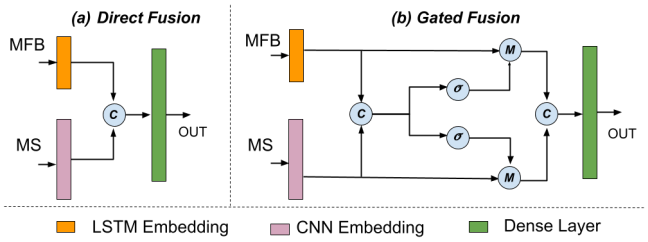

Fig. 1: The two feature embedding fusion strategies, (a) direct fusion and (b) gated fusion. The symbol C represents concatenation, σ is the Sigmoid function and M is a multiplication.

2) Estimators: The NISA+ method is designed for the joint assessment of reverberation, noise level, voice activity, codec, speech quality and intelligibility parameters. For the reverberation parameters, NISA+ estimates the C_{50}, DRR, room volume and reflection coefficient. We do not model the T_{60} parameter as this is not defined for all RIRs. Instead we estimate the room volume and reflection coefficient, which allow simulation of RIRs from reverberant speech. For the noise level, we estimate the Segmental Signal-to-Noise Ratio (SSNR) in 20 ms frames. For speech quality estimation, we use the intrusive PESQ method to label the utterances.

Model	WER	WERR-Clean	WERR-Random
Clean	47.30	-	-
Random DA	12.80	72.9%	-
NISA+ I DA	11.61	75.4%	9.2%
NISA+ II DA	11.57	75.5%	9.6%

TABLE I: ASR results for room adaptation using NISA estimated parameters, guiding data augmentation. NISA I refers to parameter selection using estimated C_{50} and SNR parameters. NISA II also uses estimated reflection coefficient and room volume parameters.

Similarly, for speech intelligibility estimation, we use the recent ESTOI [26] algorithm. In this work, we use the Opus codec [27], operating in the music and VoIP modes, at a range of bit rates from 8 to 64 kbps. Lastly, a VAD estimation is included to distinguish between speech and non-speech segments. This label is obtained by assigning each 10 ms frame to a binary class and then averaging those labels over the context window (400 ms), thus obtaining a posterior, VADP, in the range 0 to 1. Each estimation task is solved by an individual worker composed of a single fully connected output layer. There are 7 regression workers (C_{50}, DRR, SSNR, PESQ, ESTOI, VADP and Bitrate) and 1 binary classification worker (codec detection). A VADP threshold of 0.5 is used in this paper. It should be noted that, in the evaluation stage, we report the accuracy and F1 score on the detection of voice activity presence by using a threshold on the VADP posterior.

C. Training Strategy

The neural networks use a Mean Absolute Error (MAE) loss function for the regression tasks and a cross-entropy loss for the classification task. The Adam [28] optimizer is used, with an initial learning rate of 10^{-4}, that is halved every 28 epochs. The systems are trained for 140 epochs.

III. Automatic Speech Recognition

In this section we describe experiments that highlight the use of the NISA+ estimated acoustic parameters for the purpose of room adaptation of an ASR system. We use an attention-based encoder-decoder (AED) E2E ASR system that uses an encoder based on ContextNet [29] and a single layer LSTM decoder [30]. For all experiments reported here, the ASR system is trained for 90 epochs. We highlight the application of the proposed NISA+ method to extract reverberation and noise parameters from a development set of utterances recorded in a target room and then use those parameters to perform a targeted data augmentation step. We compare this approach with a baseline random augmentation of the training data. The results for the different data augmentation sets are presented in Section V based on data described in Section IV.

IV. Data and Evaluation

A. NISA+ Training Data

We follow the data preparation methodology in [11]. The training data is artificially generated by convolving clean speech from the Wall Street Journal (WSJ) corpus [31] with RIRs simulated using the Image method [32], followed by the addition of noise and finally, processing it through an Opus codec, resulting in the processed utterance, $y(t)$, as follows.

$$y(t) = F(x(t) * h(t) + n(t)),$$

where $x(t)$ is the input clean speech at discrete time instant t, $h(t)$ is the RIR, $n(t)$ is a noise source, $x(t) * h(t)$ is the convolution between $x(t)$ and $h(t)$ and F is the codec operation.

A total of 30 hours of clean speech from the WSJ corpus is used for generating the training data, which is convolved with 18,400 simulated RIRs. The C_{50} values range from 0 to 25 dB and the DRR from -15 to 13 dB. The noise sources include Ambient, White, Fan, Babble and Music and are added in a 0 to 30 dB SNR range. This is followed by the application of three codec classes (uncompressed, Opus-music, Opus-voip). An Opus codec covering bit rates in the 8 to 64 kbps range is applied in addition to uncompressed data (i.e. where no codec is applied and is set a bit rate of 128 kbps). Finally, a level augmentation in the -0.1 to -10 dBFS range is randomly applied to each utterance to account for level differences in real data.

B. ASR Training Data

The data used for training the ASR system in based on 460 hours of speech from the Librispeech [33] and Mozilla Common Voice (MCV)[1] data sets. A large set of RIRs are simulated using the Image method [32], from which 300 RIRs are selected and applied to the training data, followed by the addition of ambient noise with a uniform random distribution of SNR. The random augmented data set (Random DA) has RIRs covering a C_{50} range of 3.3 to 17.6 dB and SNRs in the 5 to 30 dB range. The NISA+ I augmented data set (NISA+ I DA) is based on selecting 300 RIRs covering a C_{50} range of 5.0 to 11.0 dB and SNRs in the 10 to 24 dB range, as estimated by NISA+. The NISA+ II augmented data set (NISA+ II DA) is based on a further selection of RIRs with a room volume range in the range 30.2 and 44.1 m^3 and reflection coefficients in the range 0.84 to 0.92, as estimated by NISA+, in addition to the C_{50} and SNR constraints as the NISA+ I DA set.

C. Test Data

In order to evaluate the NISA+ method, we use the test sets described below including a playback recorded test set for the purpose of ASR room adaptation evaluation.

1) ACE Test Set: We use a modified version of the original ACE Challenge test-set [8], by processing the original data through three codec conditions mentioned in the description of the training data. The modified ACE challenge test set contains the original (anechoic) speech material convolved with measured RIRs and additive noise recorded in the same location and rooms as the RIRs. We use this test set to evaluate the NISA+ method's performance. However, since the ACE

[1]https://commonvoice.mozilla.org/en

Model	Features	MAE						F1 Score	
		C_{50} (dB)	DRR (dB)	SSNR (dB)	PESQ Score	ESTOI Score	BitRate (kbs)	VAD	codec
CNN	MFB	2.14	3.22	4.65	0.30	**0.07**	11.26	0.92	0.89
LSTM	MFB	2.10	**3.09**	6.32	0.29	**0.07**	11.88	0.94	0.87
CNN	MS	2.34	3.36	4.07	0.28	**0.07**	7.70	0.93	**0.91**
LSTM	MS	2.44	3.40	4.88	0.31	0.08	11.89	0.92	0.90
Joint-Direct	Both	**2.01**	3.18	4.26	**0.27**	**0.07**	**7.76**	0.94	**0.91**
Joint-Gating	Both	2.04	3.24	**4.03**	**0.27**	**0.07**	7.87	**0.95**	**0.91**

TABLE III: Model evaluation of C_{50} and DRR on the ACE test dataset and other parameters on the LibriSpeech test dataset.

RIRs are measured in actual rooms, we lack the reflection coefficient labels for this data. We extract the ground truth labels using the measured RIRs and the clean speech signals, including C_{50} and DRR labels following the definition in [10].

2) Libri RIR Test Set: In order to properly test the reverberation metrics, including the room volume and reflection coefficient estimates, we create a new test set based on speech from the Libri test set [33] and simulated RIRs using the image source method [32]. The simulated RIRs allow us to cover a range of room volumes (28 to 54 m^3) and reflection coefficients (0.5 to 0.9). We note that such simulations use an idealized physical setup in which all surfaces have the same reflection coefficient at all frequencies. Nevertheless, this does allow us to perform data augmentation effectively over current approaches.

3) Libri Playback Test Set: This is a playback recording of 500 Librispeech utterances from the test-clean partition with an 8 channel uniform linear array. In this paper, we use 13 utterances from one speaker as a development set from which NISA+ parameters can be extracted and the remaining 487 utterances are used for testing the ASR systems, using channel 4 data (representing a single distant microphone). More details of this test set can be found in [34].

D. Evaluation Metrics

In the following, $P(n)_e$ and $P(n)_t$ are the estimated and true scores for a frame n of 5 ms duration. AS shown in Table II, we use the Mean Absolute Error (MAE) metric for the regression tasks and the F1 score for classification based tasks (VAD and codec presence). The ASR performance is measured in terms of the Word Error Rate (WER).

Metric	Description
Mean Absolute Error (MAE)	$\text{MAD} = \frac{1}{N}\sum_{n=1}^{N} \lvert P(n)_e - P(n)_t \rvert$
F1 Score	$\text{F1} = 2.\frac{\text{precision.recall}}{\text{precision+recall}}$
Word Error Rate (WER)	$\text{WER} = \frac{\text{Substituions+Deletions+Insertions}}{\text{Substituions+Deletions+Correct}} \times 100\%$

TABLE II: A summary of the evaluation metrics.

V. Results

A. NISA+

Table III summarizes the NISA+ results, where we can see that the single feature systems have their strengths and weaknesses in estimating different parameters. The LSTM model achieves better results with the MFB features while the CNN model works better for the MS features. The MFB features work better for C_{50}, DRR and VAD estimation while SSNR, Bitrate and codec parameters are more accurately estimated with the MS features. This suggests that a combination of the two speech features and the neural network models is required to achieve the best overall performance across all tasks. The last two rows in Table III show that the dual feature and network models outperform all single-feature and network models. Furthermore, since all models have a similar number of trainable parameters (roughly 700k), this improvement should be attributed to the fusion strategy. Given the slightly increased model complexity introduced by the gating block, the feature fusion based on direct concatenation is proposed as the best solution. The C_{50} and PESQ performance of the proposed direct feature embedding concatenation system is 16% and 15% better in relation to our previous system [11]. Table IV presents the results for the Libri RIR test set, where the estimation performance of the additional reverberation parameters (room volume and reflection coefficient) is shown. As seen, the C_{50}, DRR, Room Volume and Reflection Coefficient estimation errors are very competitive (we note that there isnt a comparable system for reflection coefficient and room volume estimation).

C_{50} (dB)	DRR (dB)	Room Volume (m^3)	Reflection Coeff.
2.16	3.05	6.43	0.12

TABLE IV: MAE results for the NISA+ reverberation parameters on the Libri RIR test-set, for the Joint-Direct NISA+ model with MFB and MS features.

B. ASR Room Adaptation

Table I presents the ASR results for different data augmentation criteria. It can be seen that the random data augmentation (Random DA) system achieves a significant reduction in WER when compared to the model trained with clean speech (72.9% WERR). The two ASR models that use NISA+ parameter estimation to perform a targeted room adaptation achieve an additional 9.2% to 9.6% WERR compared to the randomly augmented model. As has been shown in the past, the C_{50} reverberation metric is highly correlated with WER [11], and here too we note that the NISA+ I DA model, that is trained with selecting RIRs with the estimated C_{50} range performs very well, achieving a 9.2% WERR over the random augmentation model. A more detailed selection of

RIRs based on additional pruning based on estimated room volume and reflection coefficients gives only a small additional improvement of 0.4%.

VI. Conclusions

We have presented a dual feature embedding based non-intrusive speech analysis method which performs joint assessment of ten acoustic signal parameters. Our novel approach fuses spectral (MFB) and modulation (MS) information using neural representations from recurrent and convolution neural architectures. Compared to single feature and network architecture based methods, this novel approach demonstrates superior performance across all speech tasks investigated, suggesting that the MFB and MS feature embeddings provide complementary information to each other. We show that speech feature fusion based on the direct concatenation achieves the best results while maintaining a relatively small complexity. We also show how the NISA+ system can be used to perform room adaptation when training an ASR model, by a guided data augmentation step leading to a 9.6% WERR compared to a model augmented with a random selection of acoustic parameters. This highlights the applicability of our proposed method for data augmentation and analysis applications in real world scenarios.

References

[1] F. Xiong, S. Goetze, and B. T. Meyer, "Estimating room acoustic parameters for speech recognizer adaptation and combination in reverberant environments," in *In Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5522–5526.

[2] L. F. Gallardo, S. Möller, and J. Beerends, "Predicting automatic speech recognition performance over communication channels from instrumental speech quality and intelligibility scores," in *in Proc. of INTERSPEECH*, 2017, pp. 2939–2943.

[3] S. Ikram and H. Malik, "Digital audio forensics using background noise," in *In Proc. of International Conference on Multimedia and Expo*, 2010.

[4] N. Kitawaki and H. Nagabuchi, "Quality assessment of speech coding and speech synthesis systems," *IEEE Communications Magazine*, vol. 26, no. 10, pp. 36–44, 1988.

[5] M. Hu, D. Sharma, S. Doclo, M. Brookes, and P. Naylor, "Speaker change detection and speaker diarization using spatial information," in *In Proc. of ICASSP*, Brisbane, Australia, Apr. 2015.

[6] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, PUBSV, 2010.

[7] P. P. Parada, D. Sharma, T. van Waterschoot, and P. A. Naylor, "Evaluating the non-intrusive room acoustics algorithm with the ACE challenge," in *In Proc. of IWAENC*, 2015.

[8] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *In Trans. of IEEE JASLP*, vol. 24, no. 10, pp. 1681–1693, Oct. 2016.

[9] P. P. Parada, D. Sharma, and P. A. Naylor, "Non-intrusive estimation of the level of reverberation in speech," in *In Proc. of ICASSP*, 2014, pp. 4718–4722.

[10] P. P. Parada, D. Sharma, J. Lainez, D. Barreda, T. van Waterschoot, and P. A. Naylor, "A single-channel non-intrusive c_{50} estimator correlated with speech recognition performance," *In Trans. of IEEE TASSP*, vol. 24, no. 4, pp. 719–732, Apr. 2016.

[11] D. Sharma, L. Berger, C. Quillen, and P. A. Naylor, "Non intrusive estimation of speech signal parameters using a frame based machine learning approach," in *In Proc. of EUSIPCO*, Amsterdam, The Netherlands, 2020.

[12] H. Gamper and I. J. Tashev, "Blind reverberation time estimation using a convolutional neural network," in *Proc. of International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018.

[13] K. Scholz, L. Leutelt, and U. Heute, "Speech-codec detection by spectral harmonic-plus-noise decomposition," in *Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers, 2004.* IEEE, 2004, vol. 2, pp. 2295–2299.

[14] D. Sharma, U. Jost, and P. A. Naylor, "Non-intrusive bit-rate detection of coded speech," in *Proc. of EUSIPCO*, 2017, pp. 1799–1803.

[15] I.-T. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[16] R. ITU-T, "Perceptual objective listening quality assessment," *Message Sequence Charts (MSC96)*, 2011.

[17] M. H. Soni and H. A. Patil, "Novel deep autoencoder features for non-intrusive speech quality assessment," in *Proc. of EUSIPCO*, 2016, pp. 2315–2319.

[18] D. Sharma, A. O. Hogg, Y. Wang, A. Nour-Eldin, and P. A. Naylor, "Non-intrusive polqa estimation of speech quality using recurrent neural networks," in *Proc. of EUSIPCO*, 2019, pp. 1–5.

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

[20] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *In Trans. of IEEE TASSP*, vol. 28, no. 4, pp. 357–366, 1980.

[21] B. E. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *In Trans. of Speech communication*, vol. 25, no. 1-3, pp. 117–132, 1998.

[22] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS computational biology*, vol. 5, no. 3, pp. e1000302, 2009.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *In Trans. of Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] M. Hussain, M. A. Haque, et al., "Swishnet: A fast convolutional neural network for speech, music and noise classification and segmentation," *arXiv preprint arXiv:1812.00149*, 2018.

[25] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[26] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *In Trans. of IEEE TASSP*, vol. 24, no. 11, pp. 2009–2022, 2016.

[27] J. M. Valin, K. Vos, and T. B. Terriberry, "Definition of the Opus audio codec. RFC 6716," https://www.ietf.org/rfc/rfc6716.txt.

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

[29] W. Han, Z. Zhang, Y. Zhang, J. Yu, C.-C. Chiu, J. Qin, A. Gulati, R. Pang, and Y. Wu, "ContextNet: Improving Convolutional Neural Networks for Automatic Speech Recognition with Global Context," in *In Proc. of Interspeech*, Oct. 2020.

[30] R. Gong, C. Quillen, D. Sharma, A. Goderre, J. Lainez, and L. Milanovic, "Self-attention channel combinator frontend for end-to-end multichannel far-field speech recognition," in *Proc. of Interspeech*, 2021.

[31] "C.I. (WSJ1) complete," 1994.

[32] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *In Trans. of JASA*, vol. 65, no. 4, pp. 943–950, 1979.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. of ICASSP*, Brisbane, Australia, 2015, IEEE.

[34] D. Sharma, R. Gong, J. Frosburgh, S. Y. Kruchinin, P. A. Naylor, and L. Milanovic, "Spatial Processing Front-End for Distant ASR Exploiting Self-Attention Channel Combinator," 2022 (to appear in Proc. of ICASSP).