ROOM ADAPTATION OF TRAINING DATA FOR DISTANT SPEECH RECOGNITION

James Fosburgh
Research & Development
Nuance Communications Inc.
USA.
james.fosburgh@nuance.com

Dushyant Sharma
Research & Development
Nuance Communications Inc.
USA.
dushyant.sharma@nuance.com

Patrick A. Naylor
Electrical & Electronic Engineering
Imperial College London
UK.
p.naylor@imperial.ac.uk

***Abstract*—We present a novel signal processing-based approach for estimating room impulse responses for augmentation of ASR training data that is best suited to the reverberation characteristics in a particular acoustic space. Our approach estimates an impulse response of a room by using a supervised adaptive system identification algorithm to extract the relative transfer function between a speech source played through a loudspeaker and recorded by a microphone. These impulse responses can then be applied to clean speech files to create an augmented training set for an ASR system. Given the availability of a small amount of this type of playback audio for a room, we show that an ASR model trained with our data augmentation approach can provide a 19% relative reduction in word error rate compared to a system using random augmentation.**

***Index Terms*—ASR, AEC, Room Adaptation, IPNLMS**

I. Introduction

The increased deployment of Automatic Speech Recognition (ASR)-based applications has been facilitated by a number of advances in front-end signal processing, data augmentation, and deep learning architectures. Of particular interest is the use case of distant ASR in an enclosed space (i.e., a room), where non-trivial amounts of reverberation are introduced into the recorded signal. Reverberation has been shown to lead to significant degradation on ASR systems that take no measures to compensate for it [1].

One approach for solving this problem involves the use of microphone arrays and beamforming [2], [3], [4]. However, when only a single microphone channel is available, the typical ASR training approach is to rely on matched condition training data collected from rooms with similar reverberation. In a situation where we want to deploy an ASR system in a distant speaking situation with very limited in-domain data, however, this becomes more challenging. Instead, it can be highly beneficial to turn to data augmentation to make up for a lack of in-domain data. One approach for doing this is to simulate far-field data by applying reverberation and noise to clean speech. In [5], the authors evaluate the effect of this method with both real and simulated room impulse responses (RIRs). Their results show significant gains in ASR performance when using this data augmentation method over a baseline that is trained using clean speech. They additionally assert that, while real RIRs lead to better gains than simulated RIRs in the case where only reverberation is applied, this gap vanishes when also adding noise to the training data.

While this approach to data augmentation provides significant gains, it is possible that additional improvements can be made by refining the RIR set to target a specific room. In [6], the authors achieve this by using a non-intrusive signal analysis (NISA) [7] approach to estimate values for speech clarity (C50), room volume, and average reflection coefficient from playback speech, obtained by recording the output from artificial mouth loudspeakers in multiple positions in a single room. These parameters are then used to select a set of simulated impulse responses that match the target characteristics of the room. The authors show a 9.6% relative improvement in WER compared to using a random selection of RIRs.

Our approach aims to achieve better room adaptation by estimating RIRs from the playback speech directly. Assuming that playback speech for a room is available, we propose the use of an adaptive system identification algorithm, such as used for acoustic echo cancellation (AEC), for room adaptation. Using the original speech as the reference signal and the playback speech as the distant signal, we employ the improved proportionate least mean square (IPNLMS) [8] algorithm to estimate the transfer function between the speaker and the microphone, which is representative of the impulse response in the playback room between those positions. These RIRs are then used to augment training data for an ASR system.

The remainder of the paper is laid out as follows: Section II describes the data augmentation methods used, Section III presents the experiments performed, Section IV presents the results of our experiments, and Section V presents our conclusions.

II. Room Adaptation

In this section, we describe the room adaptation techniques applied in this paper. In all cases, we adopt a two-step process, in which we first obtain a set of RIRs that closely match the target room's reverberation environment, then apply the RIRs to clean training data in the second step. An ASR system is then trained on the augmented data.

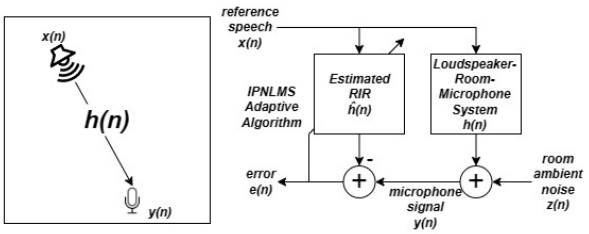


Fig. 1: Example room setup (left) and corresponding block diagram of the IPNLMS configuration for RIR estimation (right). The RIR between a source signal $x(n)$ and a microphone $y(n)$ is denoted as $h(n)$. The ambient noise component is handled by data augmentation and is not estimated our approach.

A. Baselines

In this paper, we consider two baseline data augmentation (DA) approaches, as described in more detail in the following.

1) Random DA: In the random data augmentation baseline, a large set of RIRs was simulated using the Image method [9] targeting a T60 range of [200 to 700] ms, from which 300 RIRs were selected randomly and applied to the training data. Additional augmentations were applied as described in Section III.A, with a range of [5 to 30] dB for the SNR of the noise.

2) NISA+ DA: For the NISA-based data selection baselines, RIR selection was performed according to acoustic parameters estimated using NISA+ [6]. For NISA+ I, C50 was used for the selection criteria as it was previously shown to correlate well with ASR performance [7]. RIRs were chosen from those with a C50 range of [5 to 11] dB. For the NISA+ II baseline, the estimated room volume and reflection coefficient parameters were added to the selection criterion. RIRs were chosen from those with a C50 range of [5.4 to 10.5] dB, a room volume range of [29.4 to 44.4] m^3, and a reflection coefficient range of [0.68 to 1.1]. For more information, see [6]. Additional augmentations were applied as described in Section III.A, with the SNR range being [10 to 24] dB for NISA+ I and [11 to 25] dB for NISA+ II.

B. IPNLMS RIR Estimation

The problem of estimating RIRs given some clean and playback recorded speech is akin to that of acoustic echo cancellation (AEC). In AEC, the acoustic system identification is performed by a supervised adaptive filtering algorithm. In this paper, we use the improved proportionate normalized least mean square (IPNLMS) algorithm described in [8], as it can be advantageously adjusted to take into account any sparseness in the RIRs by interpolating between the update values returned by the normalized least mean square (NLMS) algorithm detailed in [10] and the proportionate normalized least mean square (PNLMS) algorithm presented in [11].

In the formulation of the IPNLMS algorithm [8], the RIR is estimated as a finite impulse response filter, denoted as $\hat{h}(n)$ and calculated in the time domain as follows.

$$\hat{\mathbf{h}}(n) = \hat{\mathbf{h}}(n-1) + \frac{\mu \mathbf{K}(n-1)\mathbf{x}(n)e(n)}{\mathbf{x}^{\mathsf{T}}(n)\mathbf{K}(n-1)\mathbf{x}(n) + \delta_{\text{IPNLMS}}}, \quad (1)$$

where μ is the overall step-size parameter, and the error signal, $e(n)$, is defined as follows

$$e(n) = y(n) - \hat{\mathbf{h}}^{\mathsf{T}}(n-1)\mathbf{x}(n), \quad (2)$$

and

$$\mathbf{K}(n-1) = \text{diag}\{k_0(n-1,...,k_{L-1}(n-1)\}. \quad (3)$$

The reference speech signal is denoted as $x(n)$, $y(n)$ is the playback signal recorded at the microphone, $k_l(n)$ is a parameter that controls the step-size of the filter tap at position l, and δ_{IPNLMS} is the regularization parameter for the algorithm, defined as

$$\delta_{\text{IPNLMS}} = \frac{1-\alpha}{2L}\delta_{\text{NLMS}}, \quad (4)$$

where L is the length of the filter and α is the control value in the range [-1 to 1] that balances between NLMS-like performance ($\alpha = -1$) and PNLMS-like performance ($\alpha = 1$). For a complete derivation of the IPNLMS algorithm, see [8].

To estimate impulse responses using the IPNLMS algorithm, the original recorded speech is passed in as the reference signal, and the microphone signal is passed in as the distant signal. The IPNLMS algorithm iteratively estimates the filter that, when applied to the reference signal, produces the distant signal. A depiction of this setup for RIR estimation is shown in Fig. 1.

Before running the IPNLMS algorithm, the signals were band-pass filtered to 200 Hz and 7900 Hz. File pairs were then normalized to equal levels and aligned to remove the delay introduced in the playback process. An artificial delay of 30ms was then reintroduced to the microphone signal to ensure it was delayed compared to the reference to ensure causality of the estimated RIRs. Finally, an energy-based voice activity detector was run on the reference file to extract the speech regions. The IPNLMS algorithm was run only on speech regions. Additional augmentations were applied as described in Section III.A, with a range of [10 to 24] dB for the SNR of the noise.

III. Experiments

In this section, we present the training, adaptation, and test data, as well as the experiments conducted to validate the IPNLMS based approach and measure its performance against baseline approaches using an end-2-end (E2E) ASR system.

A. Training Data

The training data used in this work is taken from the LibriSpeech (LS) [12] corpus and the Mozilla Common Voice[1] (MCV) corpus. We used utterances from the 460 hours clean subset of LS training data combined with utterances from the MCV corpus, and retained a subset of 13.8 hours as a validation set. In addition to the application of simulated/estimated

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

RIRs, the following augmentations were applied to the training data: the speech level was augmented to be in the range from [-1 to -15] dBFS, white noise was added to the speech at an SNR of 45 dB to simulate microphone self-noise, and ambient noise was added at an SNR range depending on the augmentation set.

B. Test Data

The test data is based on playback of 500 utterances from the LS test-clean subset [12] through artificial mouth loudspeakers and recorded by a wall mounted 8 microphone array. The playback and recording was performed from four positions in a typical office room (3 m × 3.7 m), simulating two pairs of conversation positions. The speaker pairs were oriented toward each other as described in [13]. In this work, we use the centre microphone signal for testing.

1) Adaptation Data: From the initial 500 utterances (as described above) used for playback, 3 speakers comprising 25 utterances were excluded for use in IPNLMS data augmentation. This is an additional two speakers comprising 12 utterances more than in [6], so the ASR decode has been rerun for those models on the updated test-set. Across the 4 playback positions, this gives 100 clean-playback pairs for use in IPNLMS data augmentation.

C. Automatic Speech Recognition

To evaluate the effectiveness of the proposed method of data augmentation, we compare the results of ASR systems trained with the different data augmentation methods on the playback test-set.

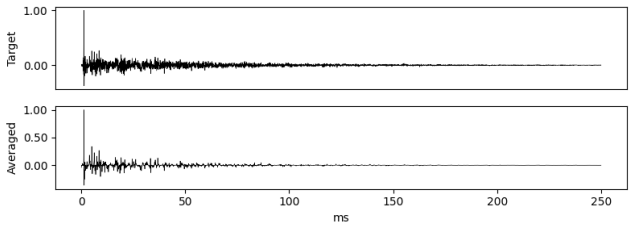
We use the same E2E automatic speech recognition system as described in [6]. The ASR system is an attention-based encoder-decoder (AED) E2E ASR system using an encoder based on ContextNet [14] and a single layer LSTM decoder [15]. For all experiments reported here, the ASR system is trained for 90 epochs, and checkpoint averaging is performed over the last 3 checkpoints to create the final model.

D. IPNLMS RIR Estimation

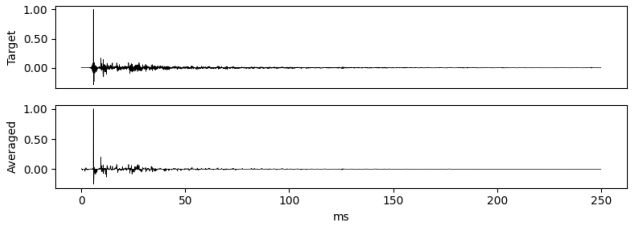
For all experiments using IPNLMS, we used an α of 0.85 and a μ of 0.1. The algorithm was run for 500k iterations, reducing μ by 5% every 10k iterations, and saving RIRs at 300k, 400k, and 500k iterations to achieve 3 RIRs per file pair.

E. Validation of IPNLMS based approach on ACE

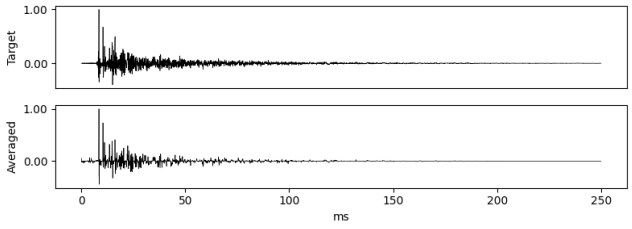
While the above experiments examine the effectiveness of using the IPNLMS algorithm through the lens of ASR performance, we performed an additional experiment using the ACE Challenge corpus [16]. This corpus provides a set of measured real RIRs, which we used to examine the accuracy of the IPNLMS estimated RIRs. For this experiment, we pulled the single channel RIRs at a distance of 2 meters from the rooms Meeting Room 1&2 and Office 1&2, and the anechoic speech files F5s3 and M6s3, which are from a female and male speaker respectively. All files were re-sampled from the original 48kHz to 16kHz. Each RIR was individually applied to both speech files, and these artificially reverberated utterances were used as the distant speech for IPNLMS while the anechoic was used as the reference. All pre-processing and settings were kept the same as described in Section II.B.1, and the final RIR was taken after 500k iterations. The resulting estimated RIR from both speakers were then averaged and compared to the real target RIR. Fig. 2 plots the target and estimated RIRs, and Fig. 3 shows the C50 and direct-to-reverberant ratio (DRR) as measured directly and as estimated by NISA+ after convolution with a speech file. As can be seen from the plots, the estimated RIRs, while more sparse than the targets, follow the overall shape of their targets well, and the trends in reverberant characteristics are additionally very well matched.



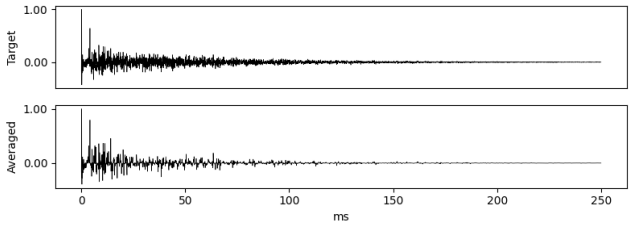
(a) Meeting Room 1



(b) Meeting Room 2

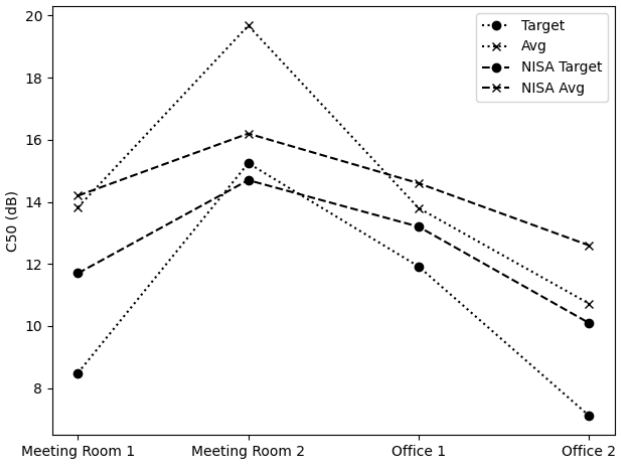


(c) Office 1

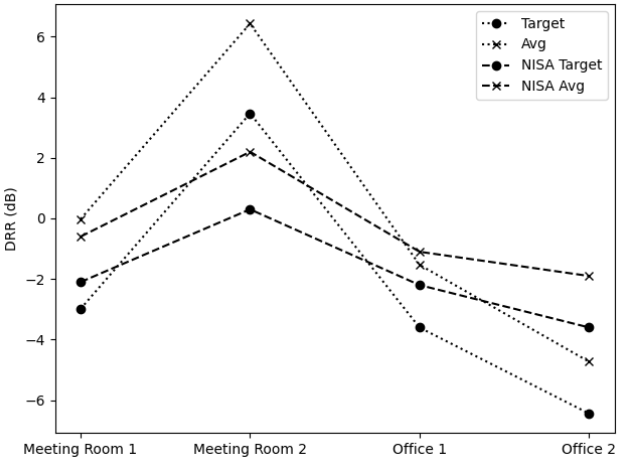


(d) Office 2

Fig. 2: Waveforms of the target real RIR and averaged IPNLMS estimated RIRs for each of the validation rooms.

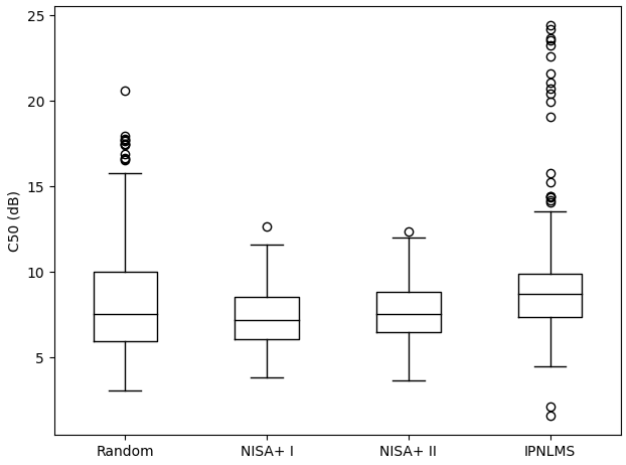


(a) C50

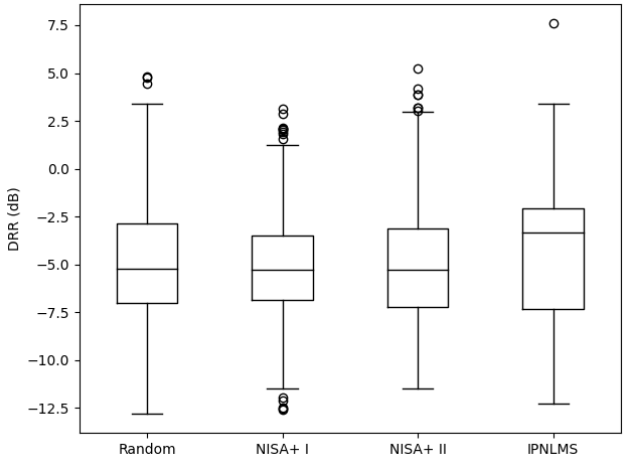


(b) DRR

Fig. 3: Plots showing the C50 (upper) and DRR (lower) of the target real RIR and averaged IPNLMS estimated RIRs as measured directly and estimated by NISA across the four validation rooms.



(a) C50



(b) DRR

Fig. 4: Comparisons of NISA+ estimated C50 (upper) and DRR (lower) between ASR training datasets augmented with the described RIR sets.

F. Metrics

As the goal of this paper is improved ASR performance, we use word error rate as our primary evaluation metric. We additionally look at the difference in distribution of C50 and DRR between the different RIR augmentation sets as measured directly from the RIRs.

IV. Results

Table I shows the ASR results for the described data augmentation (DA) techniques. As was shown in [6], all methods of data augmentation lead to greatly improved ASR performance, with a minimum WERR of 73.2% compared to the clean model. Furthermore, the models adapted specifically to a target environment show a minimum 11.2% WERR over random augmentation. Finally, using IPNLMS for data augmentation resulted in an 8.3% WERR over the best model using the NISA+ data augmentation approach.

Model	**WER**	**WERR Clean**	**WERR Random**	**WERR NISA+ II**
Clean	48.89	-	-	-
Random DA	13.08	73.2	-	-
NISA+ I DA	11.61	76.3	11.2	-
NISA+ II DA	11.57	76.3	11.5	-
IPNLMS DA	10.61	78.3	18.9	8.3

TABLE I: WER results (%) of ASR models trained with different RIR sets on LS playback test-set.

We have additionally compared the reverberant characteristics of each RIR set. Fig. 4 shows box plots of the C50 and DRR of each RIR set. In the case of C50, the 3 targeted sets have a tighter distribution than the random set, which is reflected by the ASR performance being better suited towards the playback set. However, the IPNLMS set shows a greater

difference in median C50 than any of the other sets, and additionally has far more outliers than the two pruned sets. This difference in median is even more apparent in the DRR, and the IPNLMS has a more spread out distribution than even the random set. These results, combined with the results of the ACE validation experiment, show that pruning based on reverberant characteristics provides significant improvements over not tuning at all, but suggest that estimating an RIR directly can lead to better performance despite the reverberant characteristics not being as statistically well matched.

A. Cross-Validation

To investigate the effectiveness of using IPNLMS data adaptation for unknown test configurations, we have additionally performed a cross-validation experiment. In this experiment, the RIR estimation process was the same, with an additional RIR taken after 200k steps of the IPNLMS algorithm. Four new ASR models were trained in the same manner as described in Section III.A, holding out the RIRs from one position for each model. Each model was then evaluated on the playback test files only from the room it was not trained on.

Playback Position	Random DA	NISA+ II DA	IPNLMS DA	IPNLMS Cross-Validation
P1	14.64	10.03	11.26	14.38
P2	11.82	10.69	10.01	10.21
P3	13.14	11.41	10.88	11.41
P4	12.77	11.12	10.28	11.93

TABLE II: WER results (%) by playback position comparing previously presented models against IPNLMS cross-validation, in which the results for each position are taken from an IPNLMS augmented model that has not seen that position.

Table II presents the WER results broken down over each position. As can be seen from the IPNLMS cross-validation results, in 2 positions the IPNLMS approach outperforms only the random augmentation approach on the unseen position, while in the other 2 positions there is only a slight degredation in performance, with better results only coming from the full IPNLMS approach.

V. Conclusions

We presented a method for targeted augmentation of training data to allow an ASR system to adapt to the reverberant characteristics of a specific room. Our approach uses the IPNLMS algorithm to estimate RIRs for given playback locations in a room and uses these estimated RIRs to perform acoustic augmentation of ASR training data. We show that our proposed approach outperforms the baseline adaptation techniques using non-intrusive signal analysis of the same playback data by relative 8.3% WER. This result suggests that tailoring ASR training data with playback-estimated RIRs from a room can outperform adaptation that looks only at the overall acoustic characteristics of the room, and that adaptive filter based algorithms are good candidates for this approach. We additionally show that using this method has mixed results when testing on unseen playback locations, highlighting the importance of having playback recordings representative of expected test-time locations.

For future work, we may consider frequency domain implementations of the IPNLMS and similar algorithms, as those allow for faster convergence and better robustness than time domain methods. A noise estimation block could also be added to both improve the accuracy of the ambient noise component and potentially lead to better estimation of the RIR, as the ambient noise would theoretically be removed from the error signal. We may additionally investigate a mixture of targeted adaptation techniques to make up for the varied performance on unseen positions.

References

[1] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-field automatic speech recognition," *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2020.

[2] J. Heymann, M. Bacchiani, and T. N. Sainath, "Performance of mask based statistical beamforming in a smart home scenario," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6722–6726.

[3] A. Chhetri, P. Hilmes, T. Kristjansson, W. Chu, M. Mansour, X. Li, and X. Zhang, "Multichannel audio front-end for far-field automatic speech recognition," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 1527–1531.

[4] C. Pan, J. Chen, and J. Benesty, "Microphone array beamforming with high flexible interference attenuation and noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1865–1876, 2022.

[5] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[6] G. Li, D. Sharma, and P. A. Naylor, "Non-intrusive signal analysis for room adaptation of ASR models," in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 130–134.

[7] 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.

[8] J. Benesty and S. L. Gay, "An improved PNLMS algorithm," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, vol. 2, pp. II–1881–II–1884.

[9] 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.

[10] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 1996.

[11] D. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 508–518, 2000.

[12] 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.

[13] D. Sharma, R. Gong, J. Fosburgh, S. Y. Kruchinin, P. A. Naylor, and L. Milanović, "Spatial processing front-end for distant asr exploiting self-attention channel combinator," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7997–8001.

[14] 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.

[15] 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.

[16] 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.