ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446719

DUAL-PATH MINIMUM-PHASE AND ALL-PASS DECOMPOSITION NETWORK FOR SINGLE CHANNEL SPEECH DEREVERBERATION

Xi Liu, Szu-Jui Chen, John H.L. Hansen

Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, TX, USA

ABSTRACT

With the development of deep neural networks (DNN), many DNN-based speech dereverberation approaches have been proposed to achieve significant improvement over the traditional methods. However, most deep learning-based dereverberation methods solely focus on suppressing time-frequency domain reverberations without utilizing cepstral domain features which are potentially useful for dereverberation. In this paper, we propose a dual-path neural network structure to separately process minimum-phase and all-pass components of single channel speech. First, we decompose speech signal into minimum-phase and all-pass components in cepstral domain, then Conformer embedded U-Net is used to remove reverberations of both components. Finally, we combine these two processed components together to synthesize the enhanced output. The performance of proposed method is tested on REVERB-Challenge evaluation dataset in terms of commonly used objective metrics. Experimental results demonstrate that our method outperforms other compared methods.

Index Terms— Blind speech dereverberation, cepstral analysis, minimum-phase, all-pass system

1. INTRODUCTION

The reverberation of speech caused by multiple sound reflections, can severely degrade speech quality and intelligibility in many distant-speech application scenarios such as hands-free telecommunication systems, speaker identification systems [1] and automatic speech recognition (ASR) systems. Although speech dereverberation has been studied for decades, and numerous approaches have been proposed to tackle this problem, it is still one of the most challenging task in the field of speech processing. This difficulty arises from the complexities associated with blind deconvolution and diverse acoustic spaces.

Reverberation, often referred to as convolutional noise in acoustic science, typically calls for conventional solutions such as inverse filtering or deconvolution [2] to restore clean speech. However, real-world room impulse responses (RIRs) are often unknown and exhibit non-minimum phase characteristics, rendering them unsuitable for stable inverse processing [3]. Oppenheim et al. [4] initially proposed homomorphic filtering for speech deconvolution. This can be achieved by converting the convolution operation to addition and applying peak-picking or low-pass filtering in the complex cepstral domain, since reverberation has distinct peaks in the quefrency axis. Further enhancements achieved through techniques such as temporal averaging and variable-length windowing [5]. Cepstrum-based algorithms were extensively explored during a certain period [6, 7]. However, these methods encountered difficulties in effectively addressing phase distortion issues, known to significantly impact sound quality [8]. On the other hand, RIRs in real-world environments are usually time-variant which also makes it difficult to track the echo path.

In recent years, DNN-based methods have been extensively explored to attempt to overcome these challenges due to their powerful ability to model complicated non-linearity functions. Han et al. [9] utilized multi-layer neural networks to effectively map clean speech in the time-frequency domain. Ernst et al. adopted convolutional neural network (CNN) which was previously used in image tasks to estimate the clean spectrogram and obtained promising result [10]. Similar studies [11, 12] were proposed to utilize residual connections and recurrent neural network (RNN) to boost the performance of the CNN structure. While various frequency domain DNN methods have been studied recently, the DNN-based cepstral domain method did not draw much attention. Because the cepstral domain speech has recognizable fundamental frequency features, we believe that emphasizing these features can enhance the recovery of speech harmonics. In addition, incorporating estimation of phase information has also not been fully considered in most previous studies.

In this study, we extend the concept of signal decomposition introduced in prior research [13]. We apply this idea to the task of single-channel speech dereverberation using a dual-path DNN structure. To begin, we evaluate the performance of a DNN-based real-cepstrum processing framework, highlighting its strengths and weaknesses. Next, we demonstrate how our proposed method effectively combines the advantages of both cepstral and spectral domain approaches, leading to improved performance. In the final section, we present the improvements achieved in objective evaluation metrics for speech quality and ASR.

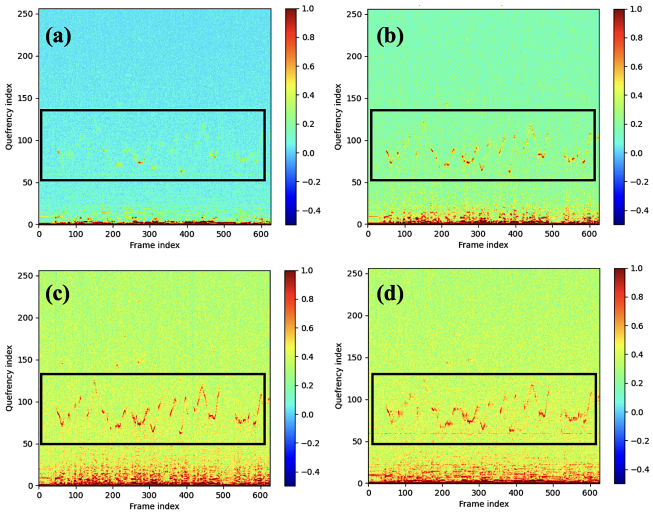

Fig. 1. Clean and reverberant speech in time-quefrency domain: (a) clean speech, (b) clean speech with α=0.5, (c) clean speech with α=0.3, (d) reverb speech with α=0.3.

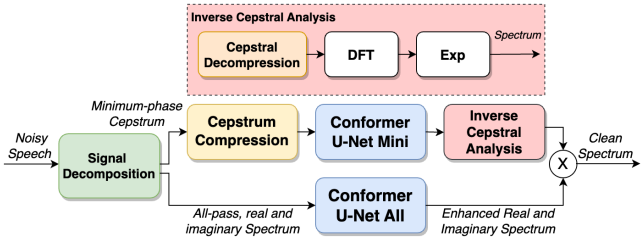

Fig. 2. The schematic diagram of proposed system.

2. PROBLEM FORMULATION

In the time domain, a single-channel speech signal corrupted with reverberation can be formulated as follows:

$$y(t) = s(t) * h(t), \tag{1}$$

where $y(t)$ is the reverberant signal captured by a omni-directional microphone, $s(t)$ is the clean speech, $h(t)$ is the room impulse response(RIR), and $*$ denotes the convolution operation. The corresponding time-frequency domain representation can be derived using the short-term Fourier transform (STFT):

$$Y(f,t) = S(f,t)H(f,t), \tag{2}$$

where f and t represent the frequency and time indices. In this study, our primary focus is on attenuating early and late reflections of speech attributed to $h(t)$. Therefore, we make the assumption that the noise level is substantially lower than the speech signal$s(t)$ and can be considered negligible.

3. PROPOSED METHOD

3.1. Cepstral analysis

Bogert [14] initially introduced the concept of cepstrum to analyze the periodicity of the spectrum. Oppenheim et al. provided a standard cepstral analysis and its inverse process [4]. By applying logarithmic operation to the two sides of Eq. 2, a deconvolution problem is converted into a subtraction in the cepstral domain. In accordance with the 'source-filter' model of speech production, speech can be decomposed into distinct components: 'excitation' and 'time-varying filter'. These components exhibit contrasting distributions in the cepstral domain, as depicted in Fig. 1. Notably, the pitch contour (shown in the black box in Fig. 1) is primarily distributed in the higher quefrency region. Likewise, the clean speech $s(t)$ in Eq. 1 is relatively distributed in the lower quefrency area compared to the distribution of the RIR $h(t)$. Based on this, it is possible to design a low-pass filter within the complex cepstral domain to preserve the speech while suppressing the reverberation. However, conventional homomorphic filtering methods encounter challenges in determining the proper cut-off queferency and solving the phase distortion problem (the clean phase is unknown).

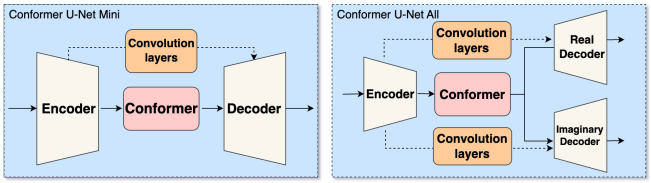

Fig. 3. U-Net structures.

3.2. DNN-based decomposition network

Instead of relying on a low-pass filter with a fixed cut-off value, we propose to estimate a universal mask for various conditions by utilizing DNN. Jiang et al. [15] proposed to separately process pitch and vocal tract of speech using cepstral analysis and DNN, although the ultimate enhancement is performed in the spectral domain.

In complex cepstral analysis, a phenomenon known as phase wrapping arises when applying a logarithmic operation to a complex value. A common solution to this issue is to unwrap the phase and wrap it back during the inverse process. This ensures that the signal can be accurately reconstructed after the cepstral analysis and its inverse. Estimating the wrapped clean phase is a major challenge, because the ground truth is typically not available. Jiang et al. [16] have attempted to solve this problem by using a complex-valued neural network. However, it could potentially introduce additional artificial noise [15].

Motivated by the study in [13], originally designed for microphone arrays, we introduce a novel approach using two parallel DNNs to individually enhance the minimum-phase and all-pass components of single-channel speech, as illustrated in Fig. 2. This concept is rooted in the factorization of the RIR into its minimum-phase and all-pass components [2]. Thus, the reverberant speech $Y(t,f)$ can be decomposed as follows:

$$Y(t,f) = Y_{mini}(t,f) \cdot Y_{all}(t,f). \tag{3}$$

$$window(n) = \begin{cases} 1 & if \quad n = 0, n = \frac{N}{2} - 1 \\ 2 & if \quad 0 < n < \frac{N}{2} - 1 \\ 0 & if \quad \frac{N}{2} - 1 < n < N - 1 \end{cases} \tag{4}$$

To be more specific, we compute the minimum-phase speech using real-valued cepstral analysis and windowing (as specified in Eq. 4, where N denotes the length of the STFT). This process avoids phase wrapping, but the real-cepstrum cannot be transformed into speech without incorporating the noisy phase. Fortunately, phase information is retained in the all-pass components, which can be easily calculated if the minimum-phase component is known (as per Eq. 3). Once both the minimum-phase and all-pass components are computed, we employ DNNs to enhance each of them separately. The reconstructed speech is then generated by multiplying the processed outputs of these two components.

$$Y_{ceptrum}^{cpr} = \begin{cases} |Y_{ceptrum}|^{\alpha} & if \quad Y_{ceptrum} \geq 0 \\ -|Y_{ceptrum}|^{\alpha} & if \quad Y_{ceptrum} < 0 \end{cases} \tag{5}$$

Furthermore, the energy of the cepstrum decreases rapidly along the quefrency axis, which result in significant lower energy of the pitch contour compared to the cepstrum in the lower quefrency region. To narrow the dynamic range and improve estimation performance, we adopt a power-law compression method to the cepstrum (Eq. 5). Where α is compression value. Fig. 1. shows time-queferency plots with different compression values. We set α to 0.3 in this study, as the corresponding plot in Fig. 1 (c) exhibits the most distinct pitch contour.

3.3. Learning target

The proposed system has two different DNNs as shown in Fig 3. The 'Conformer U-Net Mini' is responsible for estimating minimum-phase clean cepstrum, and its output yields ideal ratio mask (IRM). Alternatively, the 'Conformer U-Net All' is used to generate the real and imaginary parts of all-pass speech. It takes both the real and imaginary spectrograms of the all-pass speech as inputs and produces corresponding estimated results. Here we utilize the real and imaginary parts of clean spectrogram as learning targets. To estimate the phase and magnitude of the speech jointly, we employ a power-law compression loss function [17] that consists of:

$$L_R = \frac{1}{T \times F} \sum_{t,f} \left| |S_{real}(t,f)|^{\beta} - |\widetilde{S}_{real}(t,f)|^{\beta} \right|^2, \tag{6}$$

$$L_I = \frac{1}{T \times F} \sum_{t,f} \left| |S_{imag}(t,f)|^{\beta} - |\widetilde{S}_{imag}(t,f)|^{\beta} \right|^2, \tag{7}$$

and

$$L_{mag} = \frac{1}{T \times F} \sum_{t,f} \left| |S(t,f)|^{\beta} - |\widetilde{S}(t,f)|^{\beta} \right|^2. \tag{8}$$

Thus, the resulting overall loss function is:

$$L_{joint} = \gamma L_{mag} + L_R + L_I, \tag{9}$$

where γ is a weight factor. In this work, we set the weight factor γ and compression factor β to 5 and 0.3 respectively.

4. EXPERIMENTS

4.1. Dataset

In this study, we conduct experiments on the REVERB Challenge corpus [18], which includes real and simulated data as the training set. The simulated portion of training set is created by blending clean speech from WSJCAM0 [19] with randomly chosen RIRs from OpenSLR26 [20] and OpenSLR28. During training, the simulated data is generated on-the-fly by randomly applying RIRs to clean speech. We evaluate our proposed method using standard objective metrics on the evaluation set (*eval*) of the REVERB Challenge corpus. This *eval* set consists of 2176 simulated and 372 real recorded audio clips. In the simulated dataset, there are three different reverberation times (T60) – 0.3, 0.6, and 0.7 seconds – corresponding to reverberant rooms labeled as 1, 2, 3, respectively.

4.2. Experimental setups

The proposed framework employs a 32ms Hanning window with a 16ms skip rate for framing speech signals sampled at 16kHz. This configuration yields a DFT length of 512, and 257 frequency and quefrency bins for each frame. We use AdamW optimizer in conjunction with the 'ReduceLROnPlateau' learning rate scheduler for training. It adjusts the learning rate with a reduction factor of 0.5 and a patience of 5. The initial learning rate is set to 0.001. The Conformer embedded U-Net architecture employs a 6-layer convolutional encoder with channel numbers (32, 32, 64, 64, 96, 256). The convolutional layers have kernel sizes of (5,3) and strides of (2,1), each followed by batch normalization and PReLU activation. The decoder mirrors the structure of encoder and uses transposed convolutions. Additional convolutional layers are introduced in the skip path to establish residual connections, with kernel sizes of (3, 3) and strides of (1, 1). Conformer [21] is adopted to model both global and local information through multi-head attention and depth-wise convolution. The 2-layer Conformer configuration comprises 4 attention heads, 256 FFN dimensions, and a 15-unit kernel size for depth-wise convolution. Notably, the final layer of 'Conformer U-Net Mini' employs a Tanh activation for IRM estimation, while 'Conformer U-Net All' employs a fully connected layer as the output layer.

For ASR experiments, we train a factorized time delay neural network (TDNN-f) [22] using the Kaldi toolkit [23]. Note that we only use the clean speech as our training data and test the proposed method on the reverberant speech with a trigram language model.

4.3. Results and analysis

We evaluate our method on the REVERB challenge *eval* set. As shown in Table 1, our proposed method demonstrates superior performance compared to WPE, WRN, and Cauchi's method in both simulated and real data segments. In comparison to weighted prediction error (WPE), the proposed solution demonstrates relative improvements ranging from

Table 1. Objective Evaluations for speech quality measures on the REVERB *eval* set

	SimData												*RealData*
Metrics	CD ↓			LLR ↓			FWSegSNR ↑			SRMR ↑			SRMR↑
Room	1	2	3	1	2	3	1	2	3	1	2	3	1
Far microphone													
Unprocessed	2.67	5.21	4.96	0.38	0.75	0.84	6.68	1.04	0.24	4.58	2.97	2.73	3.19
WPE [24]	2.42	5.04	4.76	0.35	0.79	0.83	7.34	0.80	0.34	**4.87**	3.15	2.99	3.85
WRN [11]	2.43	4.99	4.56	0.35	0.59	0.67	7.54	1.79	0.88	4.48	3.32	2.84	-
Cauchi et al. [25]	2.67	4.65	4.44	0.42	0.77	0.82	8.93	3.50	2.75	4.75	3.88	3.86	4.76
Proposed	**2.22**	**4.57**	3.93	**0.28**	**0.55**	**0.60**	**10.27**	**4.92**	**4.66**	4.85	**4.76**	**4.35**	**5.79**
Cepstral branch	2.67	5.01	4.64	0.42	0.91	0.94	6.93	1.09	0.88	4.37	3.51	3.22	3.49
Spectral branch	2.56	4.57	**3.79**	0.33	0.61	0.61	8.90	3.94	4.63	3.59	3.23	3.13	4.01
Near microphone													
Unprocessed	1.99	4.63	4.38	0.35	0.49	0.65	8.12	3.35	2.27	4.50	3.74	3.57	3.17
WPE	**1.82**	4.53	4.12	0.33	0.52	0.62	8.66	3.44	2.69	4.51	3.89	3.92	4.42
WRN	2.02	4.61	4.15	0.36	0.46	0.60	8.28	3.57	2.54	4.04	3.46	3.27	-
Cauchi et al.	2.02	**3.82**	3.67	0.36	0.51	0.64	10.29	6.19	4.89	4.65	4.32	4.27	4.87
Proposed	1.89	4.11	**3.55**	**0.27**	**0.42**	**0.50**	**11.02**	**6.54**	**6.27**	**4.87**	**4.56**	**4.56**	**5.51**
Cepstral branch	2.23	4.46	4.13	0.37	0.70	0.78	8.38	2.52	2.72	4.49	3.99	3.98	3.13
Spectral branch	2.43	4.22	3.55	0.32	0.55	0.55	9.50	4.42	5.54	3.47	3.23	3.28	3.88

Table 2. WER(%)↓ performance of acoustic model trained on clean speech and test on the REVERB *eval* set.

Methods	*SimData*		*RealData*	
	Near	*Far*	*Near*	*Far*
Unprocessed	90.34	91.86	46.78	15.72
WPE	84.71	87.93	40.63	11.79
Proposed	**54.39**	**54.74**	**16.22**	**7.29**
Cepstral branch	88.93	92.24	52.68	25.99
Spectral branch	60.40	60.43	17.33	9.1

8.3% to 51.1% in cepstrum distance (CD), log likelihood ratio (LLR), and speech-to-reverberation modulation energy ratio (SRMR). On average, it achieves a 3.5dB improvement in frequency-weighted segmental signal-to-noise ratio (FWSegSNR). We conducted an ablation study to illustrate the considerable improvement achieved by the dual-branch approach compared to each individual branch. In Table 1, the 'Cepstral branch' means exclusive processing using the cepstral domain network, while the 'Spectral branch' represents results obtained solely using "Conformer U-Net All" for signal estimation. To ensure a fair comparison, we adjusted both single-branch networks to match the model size of the dual-branch network. The 'Cepstral branch' exhibits the worst performance, indicating that incorporating minimum-phase speech and noisy phase can not effectively remove reverberation. This aligns with our expectations since the all-pass response tends to have severer reverberation than the minimum-phase response [13]. The proposed method achieves better evaluation scores than single-branch methods by combining cepstrum enhancement and phase estimation. This indicates that speech dereverberation can benefit from cepstrum when clean phase estimation is incorporated, as cepstrum directly exhibit the fundamental frequency information which is useful for recovering speech harmonics. Audio samples are provided online[1].

We proceed to evaluate the effectiveness of our method by assessing its impact on an ASR system. The results in Table 2 indicate that the proposed method outperforms other approaches in various reverberation environments. Specifically, our method achieves 24.41% and 4.5% absolute improvements in WER in *near* and *far* condition of *RealData* when compared to WPE. This underscores how our method can contribute to improved performance in the back-end speech system.

5. CONCLUSIONS

In this study, we introduced a dual-path speech dereverberation system that integrates cepstral and spectral domain features by separately estimating minimum-phase and all-pass speech components. Our experimental results demonstrate that the proposed method effectively leverages the strengths of both domain features and achieves superior performance in terms of speech quality and ASR performance. In future research, we will continue to explore methods that can simultaneously suppress noise and reverberations.

[1] Audio Samples: https://winkee520.github.io/DecompNet/

6. REFERENCES

[1] Seyed Omid Sadjadi and John H. L. Hansen, "Blind spectral weighting for robust speaker identification under reverberation mismatch," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 5, pp. 937–945, 2014.

[2] Stephen Neely and Jont Allen, "Invertibility of a room impulse response," *The Journal of the Acoustical Society of America*, vol. 66, pp. 165–169, 07 1979.

[3] John Mourjopoulos, "Digital equalisation of room acoustics," *AES: Journal of the Audio Engineering Society*, vol. 42, pp. 884–893, 11 1994.

[4] A.V. Oppenheim, R.W. Schafer, and T.G. Stockham, "Nonlinear filtering of multiplied and convolved signals," *Proceedings of the IEEE*, vol. 56, no. 8, pp. 1264–1291, 1968.

[5] D. Bees, M. Blostein, and P. Kabal, "Reverberant speech enhancement using cepstral processing," in *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, 1991, pp. 977–980 vol.2.

[6] Harish Padaki, Karan Nathwani, and Rajesh M Hegde, "Single channel speech dereverberation using the lp residual cepstrum," in *2013 National Conference on Communications (NCC)*, 2013, pp. 1–5.

[7] Timo Gerkmann, "Cepstral weighting for speech dereverberation without musical noise," in *2011 19th European Signal Processing Conference*, 2011, pp. 2309–2313.

[8] B.D. Radlovic and R.A. Kennedy, "Nonminimum-phase equalization and its subjective importance in room acoustics," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 728–737, 2000.

[9] Kun Han, Yuxuan Wang, DeLiang Wang, William S. Woods, Ivo Merks, and Tao Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.

[10] Ori Ernst, Shlomo E. Chazan, Sharon Gannot, and Jacob Goldberger, "Speech dereverberation using fully convolutional networks," *arXiv preprint arXiv:1803.08243*, 2019.

[11] Dayana Ribas, Jorge Llombart, Antonio Miguel, and Luis Vicente, "Deep speech enhancement for reverberated and noisy signals using wide residual networks," 2019.

[12] Vinay Kothapally, Wei Xia, Shahram Ghorbani, John H.L. Hansen, Wei Xue, and Jing Huang, "SkipConvNet: Skip Convolutional Neural Network for Speech Dereverberation Using Optimally Smoothed Spectral Mapping," in *Proc. Interspeech 2020*, 2020, pp. 3935–3939.

[13] Qing-Guang Liu, Benoît Champagne, and Peter Kabal, "A microphone array processing technique for speech enhancement in a reverberant space," *Speech Commun.*, vol. 18, pp. 317–334, 1996.

[14] Bruce P. Bogert, "The quefrency analysis of time series for echoes : cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Proceedings of the Symposium on Time Series Analysis*, 1963, pp. 209–243.

[15] Wenbin Jiang, Tao Liu, and Kai Yu, "Efficient Speech Enhancement with Neural Homomorphic Synthesis," in *Proc. Interspeech 2022*, 2022, pp. 986–990.

[16] Wenbin Jiang, Zhijun Liu, Kai Yu, and Fei Wen, "Speech enhancement with neural homomorphic synthesis," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 376–380.

[17] Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *AAAI Conference on Artificial Intelligence*, 2019, pp. 9458–9465.

[18] Keisuke Kinoshita, Marc Delcroix, Takuya Yoshioka, Tomohiro Nakatani, Emanuel Habets, Reinhold Haeb-Umbach, Volker Leutnant, Armin Sehr, Walter Kellermann, Roland Maas, Sharon Gannot, and Bhiksha Raj, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[19] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, 1995, vol. 1, pp. 81–84 vol.1.

[20] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev 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.

[21] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented transformer for speech recognition," 10 2020, pp. 5036–5040.

[22] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks.," in *Interspeech*, 2018, pp. 3743–3747.

[23] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number CONF.

[24] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[25] Benjamin Cauchi, Ina Kodrasi, Robert Rehr, Stephan Gerlach, Ante Jukic, Timo Gerkmann, Simon Doclo, and Stefan Goetze, "Joint dereverberation and noise reduction using beamforming and a single-channel speech enhancement scheme," 05 2014, pp. 1–8.