INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Short-term Extrapolation of Speech Signals Using Recursive Neural Networks in the STFT Domain

Maurice Oberhag[1], *Daniel Neudek*[1], *Rainer Martin*[1], *Tobias Rosenkranz*[2], *and Henning Puder*[2]

[1]Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany
[2]WS Audiology, Erlangen, Germany

[1]{firstname.lastname}@rub.de [2]{firstname.lastname}@wsa.com

Abstract

This paper investigates several approaches for the short-term extrapolation of speech signals. The signal extrapolation methods are embedded into a nested two-stage spectral analysis-synthesis system for single-channel noise reduction in hearing aids. They predict additional signal samples in the low-frequency sub-bands of the first analysis stage and may compensate the additional algorithmic latency of the second, higher-resolution analysis stage in these bands. We thus achieve a higher spectral resolution in frequency bands below 3 kHz without increasing the algorithmic latency of the overall system. In the context of noise reduction, especially female voices benefit from the increased spectral resolution in the lower sub-bands of the first stage. We show that among the investigated approaches, both recursive neural-network-based extrapolation methods provide benefits in conjunction with a noise reduction algorithm and outperform our baseline linear extrapolation method.

Index Terms: extrapolation, echo-state network, low-latency, GRU, analysis-synthesis filter bank system

1. Introduction

Noise reduction is an essential component of many speech communication systems including smartphones and hearing aids. In the latter application, noise reduction algorithms aim not only at an improvement of speech quality but must also comply with strict constraints in terms of computational complexity and algorithmic latency [1, 2, 3]. The minimization of algorithmic latency is especially important for open-fitted hearing aids where the superposition of direct acoustic sounds and (delayed) amplified sounds may give rise to spectral distortions [4]. In practice, this constrains the length of input segments and thus spectral resolution of the analysis-synthesis system.

In this work we investigate several approaches for spectral analysis-synthesis that include a signal extrapolation algorithm to reduce the algorithmic latency. In particular, we use a two-stage analysis-synthesis system based on the discrete Fourier transform (DFT) and similar to [5]. There, authors have shown that an increased spectral resolution at frequencies below 3-4 kHz helps to suppress audible residual noise, especially for female voices. Therefore, after a first low-resolution filter bank stage, a second stage is implemented that provides a higher resolution in the above frequency range. Although we use a low-latency analysis-synthesis approach [6] in this second stage, some additional latency is introduced. It is the objective of the present work to eliminate this additional latency.

The extrapolation of audio signals and its combination with a noise reduction system has been considered before, most notably in [7, 8]. In these works the authors use a linear prediction approach to extrapolate signal samples and use these to extend the DFT frames used in the analysis stage. While the linear prediction approach serves as a baseline in our work, we also develop two nonlinear neural-network based methods, since acoustic speech signals are produced by a complex nonlinear physical system [9], and compare these on data sets with clean and noisy speech. The extrapolation methods are modular and detached from the noise reduction (NR) as one might choose to apply it in selected frequency bins only and a full-fledged neural network for joint extrapolation and NR might be too costly to be operated at all times.

2. System overview

Our proposed system is based on the two-stage DFT-based analysis-synthesis (AS) system introduced in [5] that uses a second AS stage for the lower frequency bands to resolve the fundamental frequencies of speech signals and, hence, enhance the NR performance, especially for female speech. To compensate the additional algorithmic latency of the second stage, we now place an extrapolator before the second stage to extend the incoming noisy signal. Fig. 1 shows a block diagram of our proposed system. The time domain input and output signals are denoted by $x(n)$ and $\tilde{x}(n)$, respectively, whereas the sub-band signal and the extrapolated samples of the sub-band signal with sub-band index μ are denoted by $x_\mu(m)$ and $\hat{x}_\mu(m)$, where m is the sub-sampled time step.

2.1. Two-stage analysis-synthesis system

Input signals are sampled at a rate of $f_{\mathrm{s},1} = 16$ kHz and processed with a standard short-time Fourier transform (STFT) in both AS stages. The first stage uses a DFT length $K_1 = 128$, a frame advance $R_1 = 32$ and periodic square-root Hann windows. The effective window length of the analysis and synthesis window are set to $L_{\mathrm{ana},1} = L_{\mathrm{syn},1} = 127$, resulting in a band distance of $\Delta f_1 = 125$ Hz, a sub-band sampling rate of $f_{\mathrm{s},2} = 500$ Hz, and an algorithmic latency of $\tau_{\mathrm{d},1} = 7.875$ ms. The second stage is applied to the lower sub-band signals up to the sub-band with sub-band index $\mu' = 23$ that corresponds to a center frequency of 2875 Hz. We set $K_2 = 16$ and $R_2 = 1$ to achieve a band distance of $\Delta f_2 = 31.25$ Hz within the second stage. To reduce the algorithmic latency of the second stage, we use asymmetric analysis and synthesis windows that are designed as proposed in [6], with effective window lengths of $L_{\mathrm{ana},2} = 15$ and $L_{\mathrm{syn},2} = 3$, resulting in an additional latency of $\tau_{\mathrm{d},2} = 4$ ms. However, in our experiments described in Section 3.2.1, we also investigate the performance of the system with $L_{\mathrm{syn},2} = 5$ and $L_{\mathrm{syn},2} = 7$. In general, the algorithmic latency

$$\tau_{\mathrm{d}} = (L_{\mathrm{syn}} - 1)\,/f_s, \tag{1}$$

10.21437/Interspeech.2023-992

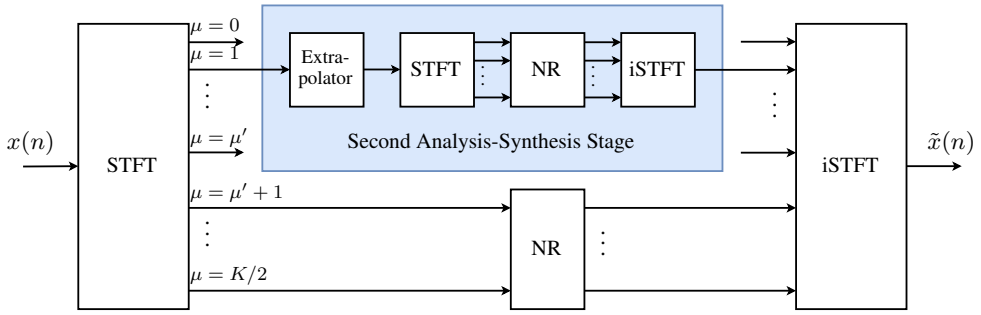

Figure 1: *Overview of our proposed two-stage analysis-synthesis system with extrapolation and noise reduction (NR). The input $x(n)$ is decomposed into K frequency sub-band signals by the first STFT. To compensate the latency of the second stage, each of the lower sub-band signals up to sub-band index μ' are extrapolated and then decomposed by a second STFT. The NR is applied within the second stage as well as on the upper sub-band signals of the first stage. Finally, the last inverse short-time Fourier transform (iSTFT) reconstructs the output signal $\tilde{x}(n)$.*

is determined by the effective synthesis window length L_{syn}, where we neglect any buffering or processing time and consider only the latency induced by the overlap-add synthesis.

2.2. Extrapolation methods

We evaluate and compare several linear and nonlinear extrapolation methods within our two-stage AS system. The linear method uses an auto-regressive (AR) model to predict future samples for each sub-band signal, separately. The AR model with N_{FO} filter coefficients extrapolates the sample $\hat{x}_\mu(m+1)$ via a linear combination of the last known samples

$$\hat{x}_\mu(m+1) = \sum_{i=0}^{N_{\text{FO}}-1} a(i) x_\mu(m-i). \tag{2}$$

We update the filter coefficients at each time step m with Burg's method [10] using an input vector $\mathbf{x}_\mu(m)$ that contains the past N_{p} samples of the respective sub-band signal

$$\mathbf{x}_\mu(m) = [x_\mu(m), \ldots, x_\mu(m - N_{\text{p}} + 1)]^{\text{T}}. \tag{3}$$

N_{FO} and N_{p} are optimized for each sub-band using a model-based global optimization (MBO) method [11] on a set of 40 clean speech signals (20 female, 20 male) of the LibriSpeech corpus [12]. Since LibriSpeech contains compressed speech samples, we select files with a crest factor $\geq$ 18 dB and a signal duration $\geq$ 15 s. The optimized parameters range in $N_{\text{FO,opt}} \in \{5, \ldots, 12\}$ and $N_{\text{p}} \in \{14, \ldots, 30\}$.

As nonlinear methods, we use two different recursive neural networks. The first approach comprises an *echo-state-network* (ESN) including a recursive least-squares (RLS) adaptive filter, as described in [13]. The ESN features a feedback network based on a sparse feedback matrix and a nonlinear activation function that generates diverse signal components which are then used to approximate each new signal sample via the adaptive filter. The method has shown excellent results in previous linear prediction tasks [14] and does not require offline training. Similar to the linear AR method, we use a separate ESN for each sub-band signal. The ESN takes an input vector $\mathbf{x}_\mu(m)$, as described in (3), to compute the nonlinear reservoir state vector $\mathbf{y}(m)$ by

$$\mathbf{y}(m) = f_{\text{a}}\left(g \cdot \mathbf{W}_{\text{in}} \mathbf{x}_\mu(m) + \mathbf{W}\mathbf{y}(m-1)\right), \tag{4}$$

where f_{a} is a nonlinear activation (hyperbolic tangent, tanh) function, g is a constant gain factor, $\mathbf{W}_{\text{in}}$ is the $M \times N_{\text{p}}$ input weight matrix that connects the input vector with M reservoir neurons, $\mathbf{W}$ is the $M \times M$ feedback matrix that connects the M reservoir neurons of the ESN via a delay unit with each other. The input weight matrix $\mathbf{W}_{\text{in}}$ contains uniformly distributed random values between -1 and 1, whereas the sparse feedback matrix uses only 10% of the possible neuron connections and contains uniformly distributed random values between 0 and 1. In addition, the spectral radius, which is the maximum of all eigenvalues, of $\mathbf{W}$ is set to 0.5. Since the input samples are complex-valued, we modify the computation of the nonlinear activation function such that we apply the activation function only to the absolute value and recombine the output with the corresponding phase. The output of the ESN is then computed by an adaptive filter $\mathbf{w}_{\text{out}}(m)$ with

$$\hat{x}_\mu(m+1) = \mathbf{w}_{\text{out}}^{\text{T}}(m) \overline{\mathbf{y}}(m), \tag{5}$$

where $\overline{\mathbf{y}}(m)$ is a concatenated vector containing the reservoir state vector and the input vector

$$\overline{\mathbf{y}}(m) = \left[g \cdot \mathbf{x}_\mu^{\text{T}}(m), \mathbf{y}^{\text{T}}(m)\right]^{\text{T}}, \tag{6}$$

and $\mathbf{w}_{\text{out}}(m)$ is adapted via the RLS algorithm. We optimize the following parameters of the ESN including the RLS algorithm on the same signals and with the same method as the AR filter and denote the resulting method as NN-ESN: the number of neurons M, the number of input samples N_{p}, the RLS regularization parameter Δ, the RLS forgetting factor λ and the constant gain g. The optimized parameters range in $M \in \{10, \ldots, 22\}$ and $N_{\text{p}} \in \{5, \ldots, 15\}$ and lead to an average adaptive filter length of 17 per sub-band. The RLS tuning parameters and the parameter of the RLS transition matrix, described in [14], are set to $\beta = 0.3$, $q = 0.25$ and $\alpha = 0.999$.

The second nonlinear method is a *pre-trained neural network* (NN-GRU) with three layers, as illustrated in Fig. 2, and 111279 parameters, in total. The network takes the real and imaginary (RI) components of the most recent sample of all lower sub-band signals and predicts the RI components of the next future sample of each sub-band. The stacked real and imaginary parts result in an input and output vector dimension of 47. The input vector is fed to a fully-connected (FC) layer with 128 neurons and tanh activation followed by a gated recurrent unit (GRU) [15] with 128 neurons and a FC layer with 47 neurons and tanh activation. The predicted future sample of each of the lower sub-bands can then be obtained by recombining the RI components from the output of the network. The

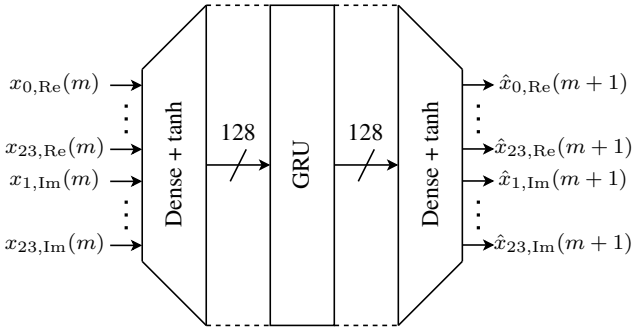

Figure 2: *Structure of the NN-GRU extrapolator with an input and output dimension of* 47 *and a GRU layer.*

network is trained on the LibriSpeech 100h dataset [12] using the ADAM optimizer [16] and the standard mean-squared error (MSE) loss function. We choose a mini-batch size of 128 and a learning rate $\lambda = 4 \cdot 10^{-4}$ that is reduced by 10% every 8 epochs. We create the input training data by processing randomly chosen 3 s snippets of each training signal with the first analysis-synthesis stage. The target data equals the input sub-band signals shifted by one sample into the future, so that the network learns to predict the next upcoming sample of the sub-band signals ($x_\mu(m) \;\rightarrow\; \hat{x}_\mu(m+1)$). We evaluate the network in each epoch using the development set of the LibriSpeech corpus to observe the training progress.

To compare the complexity of the nonlinear extrapolators to the linear method, we measure the execution time of all extrapolation methods in the proposed system with $L_{\mathrm{syn},2} = 3$. While the NN-ESN requires only 60% of the processing time of the AR method, the NN-GRU takes approximately twice as long as the AR method.

3. Evaluation

We evaluate the quality of the extrapolation process and the overall performance of the proposed system in terms of the short-time objective intelligibility (STOI) measure [17] and the perceptual evaluation of speech quality (PESQ) measure [18]. Moreover, for our experiments with clean speech signals, described in Section 3.1, we also compute the prediction gain

$$\mathrm{G_p}(\mu) = 10\log_{10}\left(\frac{\mathrm{var}\{x_\mu(m)\}}{\mathrm{var}\{x_\mu(m) - \hat{x}_\mu(m)\}}\right), \qquad (7)$$

for each sub-band signal μ, where $\mathrm{var}\{.\}$ computes the variance of its input, and for the reconstructed time domain signal

$$\mathrm{G_p} = 10\log_{10}\left(\frac{\mathrm{var}\{x(m)\}}{\mathrm{var}\{x(m) - \tilde{x}(m)\}}\right). \qquad (8)$$

For the experiments with noisy speech signals, described in Section 3.2, we compute the output signal-to-noise ratio

$$\mathrm{SNR_{out}} = 10\log_{10}\left(\frac{\mathrm{var}\{s(m)\}}{\mathrm{var}\{s(m) - \tilde{x}(m)\}}\right), \qquad (9)$$

where $\tilde{x}(m)$ is the reconstructed (enhanced) time domain signal and $s(m)$ is the respective clean speech signal.

3.1. Extrapolation of clean speech signals

In an initial experiment, we extrapolate 16 clean speech signals (8 male, 8 female) of the LibriSpeech test set [12] with a crest factor $\geq$ 18 dB and a signal duration $\geq$ 15 s using different extrapolation lengths. For this purpose, we only use the first analysis-synthesis stage, extrapolate the lower sub-band signals and synthesize the time domain signal using the predicted sub-band samples. To evaluate the effect of the extrapolation more precisely, we set the higher sub-band signals to zero before reconstructing the time domain signal.

Fig. 3a-3c show the prediction gain for different extrapolation lengths evaluated on three sub-band signals. The prediction gains for the remaining sub-band signals are similar to the trend that is shown in the figure. The prediction gain declines rapidly for larger extrapolation lengths. Moreover, the extrapolation achieves higher prediction gains in lower frequency bands. As a result, the extrapolation length should be set ≤ 4 ms to minimize distortions. The NN-GRU shows promising performance for a short extrapolation length below or equal to 4 ms. However, this comes at the cost of a higher computational complexity. On the other hand, NN-ESN achieves a similar performance as the AR method for the aforementioned extrapolation lengths at a lower computational complexity.

We also evaluate the reconstructed time domain signal, see Fig. 3d-3f. To remove the higher frequency bands, the reference signal that is used for the evaluation metrics is processed by the same AS system but without extrapolation. For PESQ and STOI the NN-GRU approach shows a more rapid degradation when the extrapolation length is increased. Focusing at extrapolation lengths ≤ 4 ms, the NN-GRU outperforms the other methods with regard to STOI and prediction gain, though it clearly provides lower PESQ values. We found that the joint extrapolation of the sub-band signals used by the NN-GRU produces a tonal artifact in the reconstructed signal with fundamental frequency of $f_{\mathrm{s},1}/R_1 = 500$ Hz and its harmonics. It is faintly audible but leads to a lower PESQ value.

3.2. Noise reduction performance

In this section we evaluate the performance of the proposed system with an algorithmic latency of 7.875 ms (see Fig. 1) and compare it to two reference systems. As an upper reference (UpR), we choose a two-stage analysis-synthesis system with the same parameters, but without extrapolation, resulting in an algorithmic latency of 11.875 ms. The lower reference (LowR) system uses a one-stage analysis-synthesis system with the same parameters as the first stage of our proposed system and, therefore, has the same algorithmic latency as our proposed system. To evaluate the contribution of the extrapolation in a transparent and reproducible fashion, we use a standard (white-box) frequency-based Wiener filter [19, 20] and the recursive noise power estimation approach [21] for all three systems. We evaluate the systems using the same test signals, as described in Section 3.1, mixed with either babble or white Gaussian noise.

3.2.1. Second-stage synthesis window length

First, we compare different synthesis window lengths $L_{\mathrm{syn},2}$ for the second stage using input signals mixed with white Gaussian noise. Since the synthesis window length also affects the shape of the analysis window and, hence, the frequency resolution, we aim to find the best trade-off between frequency resolution of the second stage and distortions caused by the extrapolation. As seen in Table 1, using a shorter synthesis window provides equal or better system performance in all metrics.

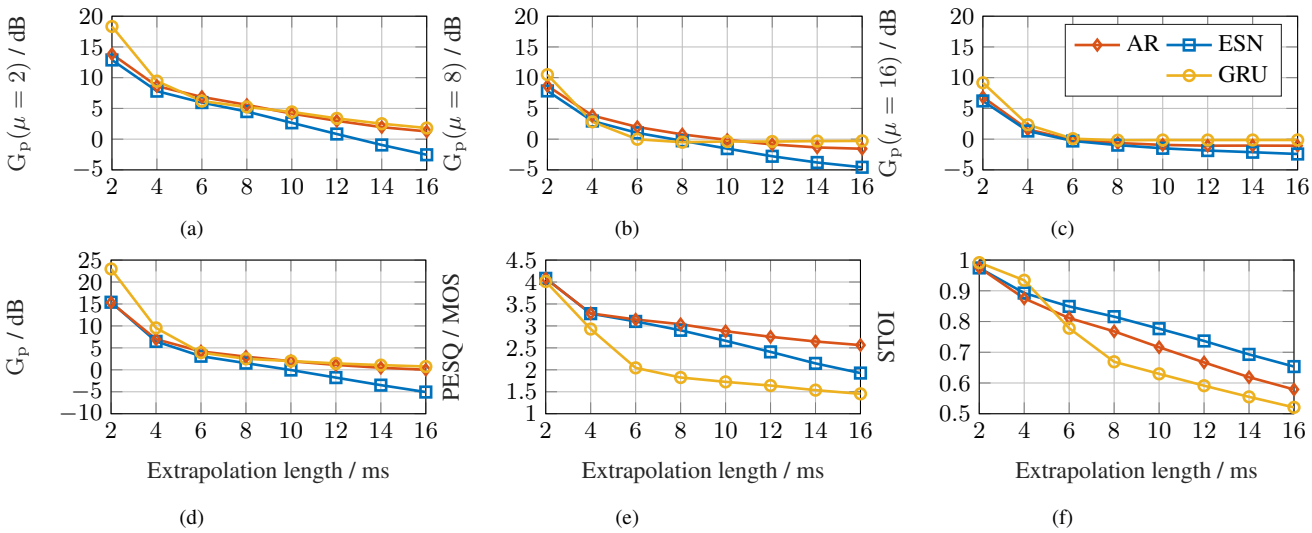

Figure 3: *Prediction gain for different extrapolation lengths evaluated on the sub-band signals with center frequency a) $f_c = 250$ Hz, b) $f_c = 1000$ Hz and c) $f_c = 2000$ Hz. d) Prediction gain, e) PESQ and f) STOI evaluated on the reconstructed time domain signal.*

Table 1: *NR performance for different second-stage synthesis window lengths $L_{\text{syn},2}$ averaged over all test signals.*

	$L_{syn,2}$	AR	ESN	GRU	UpR	LowR	Input
SNR_{out}	3	8.69	8.88	**9.49**	10.15	10.10	5
	5	6.26	6.12	7.07	10.32	10.10	
	7	4.61	4.11	5.00	10.44	10.10	
PESQ	3	2.11	2.15	2.08	2.11	2.07	1.80
	5	2.09	**2.16**	2.02	2.12	2.07	
	7	2.05	**2.16**	1.92	2.13	2.07	
STOI	3	80.56	81.07	**81.97**	82.31	81.68	81.08
	5	76.36	77.93	80.38	82.45	81.68	
	7	72.53	75.01	78.01	82.54	81.68	

3.2.2. *SNR and gender dependent evaluation*

On the basis of the results, described in 3.2.1, we test the system at different SNR using the short synthesis window with $L_{\text{syn},2} = 3$ and input signals mixed with babble noise. Moreover, we average the results over female and male speakers, separately, as shown in Table 2. As we can see, the two-stage system provides a clear benefit for the NR performance in comparison to the one-stage system, especially for female speech signals. Moreover, both of our nonlinear neural-network-based methods clearly outperform the linear baseline method. For STOI and SNR_{out} the NN-GRU achieves the best performance among the extrapolation methods and is close to the upper reference system in terms of PESQ and STOI. By contrast, the NN-ESN achieves promising PESQ improvements up to 0.14 MOS w.r.t. the upper reference system for female speech signals.

4. Conclusions and outlook

We presented different methods for the short-term extrapolation of speech signals and used them in a two-stage analysis-synthesis system to compensate the algorithmic latency of the second filter bank stage. Among these methods, our proposed GRU-based neural network achieves the best performance on extrapolating clean speech using an extrapolation length of 2 ms or 4 ms. Moreover, both our NN-based systems achieve promising benefits for the noise reduction, especially visible for female speech, and outperform our linear baseline system. Our proposed methods are easily integrable with a low computational complexity into existing noise reduction systems, e.g. in hearing aids, to improve noise reduction performance without increasing the algorithmic latency. However, the joint extrapolation of the sub-band signals used by our NN-GRU leads to a faintly audible disturbing tone that rapidly degrades the PESQ value for larger extrapolation lengths. Thus, low-complexity networks for extended extrapolation will be the main focus of future work.

Table 2: *NR performance for different SNR.*

(a) *Averaged over **female** speakers.*

	SNR	AR	ESN	GRU	UpR	LowR	Input
SNR_{out}	0 dB	4.45	4.40	**4.57**	4.25	3.97	0
	5 dB	7.50	7.90	**8.26**	8.39	7.87	5
	10 dB	10.33	10.51	**11.42**	12.36	11.80	10
PESQ	0 dB	1.68	**1.71**	1.65	1.66	1.60	1.49
	5 dB	2.09	**2.16**	2.03	2.02	1.96	1.80
	10 dB	2.47	**2.57**	2.43	2.41	2.36	2.14
STOI	0 dB	63.86	64.32	**65.49**	65.44	64.02	64.35
	5 dB	76.39	**77.10**	76.89	77.00	75.60	76.15
	10 dB	84.86	85.74	**86.02**	86.27	85.24	85.90

(b) *Averaged over **male** speakers.*

	SNR	AR	ESN	GRU	UpR	LowR	Input
SNR_{out}	0 dB	3.95	**4.25**	4.15	3.89	3.71	0
	5 dB	7.05	7.50	**7.93**	8.21	7.88	5
	10 dB	9.50	10.06	**11.05**	12.17	11.97	10
PESQ	0 dB	1.69	**1.73**	**1.73**	1.74	1.70	1.62
	5 dB	2.03	**2.08**	2.05	2.04	2.03	1.85
	10 dB	2.40	**2.46**	2.42	2.42	2.41	2.19
STOI	0 dB	64.96	64.76	**66.14**	66.33	65.38	65.58
	5 dB	75.10	75.63	**76.92**	77.21	76.62	76.89
	10 dB	83.34	83.81	**85.48**	85.78	85.64	85.55

5. References

[1] H. Löllmann and P. Vary, "Low delay noise reduction and dereverberation for hearing aid," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 1 – 9, 2009.

[2] H. Schröter, T. Rosenkranz, A.-N. Escalante-B, and A. Maier, "Low latency speech enhancement for hearing aids using deep filtering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2716–2728, 2022.

[3] P. Vary, "An adaptive filterbank equalizer for speech enhancement," *Signal Processing*, vol. 86, pp. 1206–1214, 2006.

[4] G. Stiefenhofer, "Hearing aid delay in open-fit devices – coloration-pitch discrimination in normal-hearing and hearing-impaired," *International Journal of Audiology*, pp. 1–9, 2022.

[5] A. Schasse, T. Gerkmann, R. Martin, W. Sorgel, T. Pilgrim, and H. Puder, "Two-stage filter-bank system for improved single-channel noise reduction in hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 383–393, 2015.

[6] D. Mauler and R. Martin, "A low delay, variable resolution, perfect reconstruction spectral analysis-synthesis system for speech enhancement," in *2007 15th European Signal Processing Conference*, 2007, pp. 222–226.

[7] I. Kauppinen and K. Roth, "Audio signal extrapolation - theory and applications," in *Proc. of the 5th Int. Conference on Digital Audio Effects (DAFx-02)*, 2002, pp. 105 – 110.

[8] ——, "Improved noise reduction in audio signals using spectral resolution enhancement with time-domain signal extrapolation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1210–1216, 2005.

[9] N. Tishby, "A dynamical systems approach to speech processing," in *International Conference on Acoustics, Speech, and Signal Processing*, 1990, pp. 365–368 vol.1.

[10] J. P. Burg, "Maximum entropy spectral analysis," dissertation, Stanford, 1975. [Online]. Available: http://sepwww.stanford.edu/data/media/public/oldreports/sep06/

[11] A. I. J. Forrester, A. Sóbester, and A. J. Keane, *Engineering Design via Surrogate Modelling*. Chichester, U.K.: Wiley, 2008.

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 5206–5210.

[13] M. Lukoševičius, "A practical guide to applying echo state networks," in *Neural Networks: Tricks of the Trade: Second Edition*, G. Montavon, G. B. Orr, and K.-R. Müller, Eds. Berlin, Heidelberg: Springer, 2012, pp. 659–686.

[14] Z. Zhao, H. Liu, and T. Fingscheidt, "Nonlinear prediction of speech by echo state networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. Rome: IEEE, Sep. 2018, pp. 2085–2089.

[15] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling." arXiv, 2014, https://arxiv.org/abs/1412.3555.

[16] D. Kingma and J. Ba, "Adam: A method for stochastic optimization." arXiv, 2014, https://arxiv.org/abs/1412.6980.

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

[18] 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 *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. Salt Lake City, UT, USA: IEEE, 2001, pp. 749–752.

[19] R. Martin, "Statistical methods for the enhancement of noisy speech," in *Speech Enhancement*. Berlin, Heidelberg: Springer, 2005, pp. 43–65.

[20] P. Vary and R. Martin, *Digital Speech Transmission-Enhancement, Coding and Error Concealment*. Chichester, U.K.: Wiley, 2006.

[21] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 145–148.