A Two-Stage Approach for Noisy-Reverberant Speech Intelligibility Improvement

G. Zucatelli, F. Farias and R. Coelho

***Abstract*— In this paper, a two-stage time domain technique is proposed to improve intelligibility of speech signals under noisy-reverberant conditions. In this method, the NNESE and $\mathrm{ARA_{NSD}}$ methods are jointly taken into account to mitigate the effects of noise and reverberation separately. Additionally, the resulting approach is adaptive in the sense that no prior knowledge of speech statistics or room information is required. Two intelligibility measures ($\mathrm{ASII_{ST}}$ and ESII) are used for objective evaluation. The results show that the proposed two-stage scheme leads to a higher intelligibility improvement when compared to competing methods, specially for low SNR values. Furthermore, the PESQ and the updated version of the SRMR quality measure ($\mathrm{SRMR_{norm}}$) demonstrate that the proposed technique also attains quality improvement.**

***Keywords*— speech intelligibility, noisy-reverberant, non-stationarity, adaptive methods**

I. Introduction

Reverberation is an acoustic effect that regularly occurs in enclosed and urban environments such as concert halls, parks and offices. This condition changes characteristics of speech and can cause quality and intelligibility reduction [1][2]. Moreover, speech signals can also be degraded by background acoustic noises (Babble and Cafeteria) present in the urban space. Such non-stationary effects are a major drawback to speech intelligibility improvement.

In the literature, speech enhancement solutions as the NNESE [3], EMDH [4] and UMMSE [5] were designed to cope with background non-stationary acoustic noises [6]. These methods rely on the estimation of noise statistics and subsequent enhancement of corrupted speech, attaining interesting results for both quality and intelligibility. However, room reverberation is not considered by these techniques.

More recently, approaches as the single-channel online enhancement (SCOE) [7], the adaptive reverberation absorption with non-stationary detection ($\mathrm{ARA_{NSD}}$) [8] and the reverberant speech enhancement (RSE) [9] account for reverberation masking effects. The first one adopts a Bayesian filtering formulation of the noisy-reverberant problem considering a trained hidden Markov model (HMM) for speech modelling. On the other hand, the $\mathrm{ARA_{NSD}}$ detects variations on the natural non-stationarity behavior of speech signals in order to preserve important speech regions. This method works similar to a physical element, changing the low absorption characteristic of materials that compose a room and mitigating the reverberation effect. At last, the RSE approach combines a dereverberation step followed by a spectral subtraction, requiring prior knowledge of room information.

In this work, a two-stage technique based on the NNESE and $\mathrm{ARA_{NSD}}$ methods is proposed for noisy-reverberant speech intelligibility improvement. The main idea is to process each distortion present on a noisy-reverberant environment separately, in two different stages. The NNESE is considered for it is designed to deal with non-stationary noises in the time-domain. Furthermore, the $\mathrm{ARA_{NSD}}$ is adopted because of its interesting results on mitigating masking effects of reverberation. A new energy normalization procedure is included on the NNESE signal reconstruction step. Both methods adaptively mitigate noise and reverberation distortions, leading to speech intelligibility and quality improvement. No prior knowledge of the room acoustics or speech statistics is required, which reinforces the adaptability of the proposed technique.

Extensive experiments are conducted to objectively evaluate the proposed approach improvements on speech intelligibility and quality. The noisy-reverberant scenario is composed of three real reverberant rooms (Meeting, Stairway and LASP1) selected from the AIR [10] and LASP_RIR[1] databases and two background non-stationary acoustic noises (Babble and Cafeteria) with SNRs of -2 dB, 0 dB and 2 dB. The $\mathrm{ASII_{ST}}$ [11] and ESII [12] objective measures are adopted for the intelligibility prediction. These measures are explicitly designed to deal with the non-stationarity of speech and noise-reverberant distortions. The PESQ is selected for quality evaluation. The $\mathrm{SRMR_{norm}}$ [13] quality measure is further considered as it is primarily used for signals under reverberation.

This paper is organized as follows. The proposed method is presented in Section II. The experiments are demonstrated in Section III followed by the Conclusion in Section IV.

II. NNESE+ARA: A Two-Stage Technique for Noisy-Reverberant Speech Signals

The proposed method is here presented considering the stages for attenuation of noise and reverberation masking effects. The first stage follows the steps of the NNESE [3]. A new normalization procedure is introduced in the signal reconstruction step of [3]. The second stage is the adaptive absorption of the $\mathrm{ARA_{NSD}}$ [8] dedicated to mitigate masking distortions, such as reverberation. A new set of sigmoid functions presented in [8] are implemented to better combine both approaches. The new technique is named NNESE+ARA.

This work was partially supported by the National Council for Scientific and Technological Development (CNPq) 308155/2019-0 and Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) 203075/2016. This work is also supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Grant Code 001. The authors are with the Laboratory of Acoustic Signal Processing (lasp.ime.eb.br), Military Institute of Engineering (IME), Rio de Janeiro, Brazil (e-mail: zucatelli, felipe.farias, coelho@ime.eb.br).

[1] Available at lasp.ime.eb.br

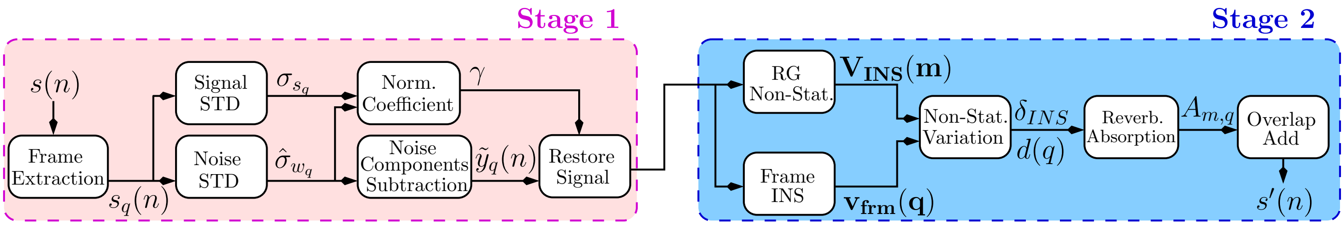


Fig. 1. Schematic overview of the proposed NNESE+ARA technique illustrated in stages 1 and 2, respectively.

The main goal is to improve speech intelligibility and quality under noisy-reverberant conditions by treating each distortion separately using the specific methods (NNESE and ARA_{NSD}) to diminish noise and reverberation, respectively.

The reverberation effect can be defined as a linear filtering process such that, given a room impulse response (RIR) $h(n)$, the reverberated signal can be obtained by convolution. The RIR is typically characterized by the reverberation time (T_{60}) and the direct-to-reverberant ratio (DRR). These parameters describe the reverberation duration until a 60 dB power reduction and the intensity relative to the direct signal, respectively. In real environments, acoustic noises are also a common distortion, which means that the resultant noisy-reverberant speech signal $s(n)$ can be obtained by

$$s(n) = x(n) * h(n) + w(n), \tag{1}$$

where $x(n)$ is the clean speech signal and $w(n)$ is the background noise. Note that, by this model, the environmental noise is additive to the reverberated signal. Therefore, it is desired to treat this noise distortion first and latter process the reverberation masking effect. To this end, the proposed technique is organized in two main stages depicted in Fig. 1.

A. Stage 1

A speech enhancement based on the NNESE technique is considered in the first stage to deal with background acoustic noises. This method can be segregated in three steps:

- noise standard deviation estimation ($\hat{\sigma}_w$) using the short-time version of the d-Dimensional Trimmed Estimator (DATE) [14] for noisy signal in the time-domain,
- selection of noise only amplitude components based on a threshold (b_w) derived from $\hat{\sigma}_w$,
- frame based speech signal reconstruction.

A normalization coefficient γ is here proposed in the speech signal reconstruction step. Given the q-th frame with $n = 1, \ldots, N$ samples, the resulting frame is calculated by

$$\tilde{y}_q(n) = \begin{cases} s_q(n) - \alpha\hat{\sigma}_{w_q}, \text{ if } y_q(n) > y(b_{w_q}); \\ \beta s_q(n), \text{ otherwise }, \end{cases} \tag{2}$$

where α and β are estimation parameters of $\hat{\sigma}_w$. After this noise attenuation the frame energy is normalized multiplying its amplitudes by a normalization coefficient γ given by

$$\gamma = \sqrt{(\sigma^2_{s_q} - \hat{\sigma}^2_{w_q})/\sigma^2_{\tilde{y}_q}}. \tag{3}$$

This way the final frame energy is guaranteed to be the difference between the signal and estimated noise energies.

B. Stage 2

The second stage of NNESE+ARA is based on the ARA_{NSD} [8] and accounts for mitigating reverberation. This is accomplished in two steps: reverberation detection and acoustic absorption. For the detection, a reverberation group (RG) is defined as the m-th segment composed of eight consecutive frames of the corrupted speech. This window duration is selected to enable a long-term temporal observation of the reverberation effect using the Index of Non-Stationarity (INS) [15]. Consecutive INS vectors are used to compute a normalized variation of the non-stationary property as

$$\delta_{INS}(m) = \frac{||\mathbf{v_{INS}}(m) - \mathbf{v_{INS}}(m-1)||}{||\mathbf{v_{INS}}(m)|| + ||\mathbf{v_{INS}}(m-1)||}. \tag{4}$$

It is demonstrated in [8] that $\delta_{INS}(m)$ can identify important intelligibility speech regions. A similar distance $d \in [0, 1]$ is computed on a frame-by-frame basis and is adopted in the frame absorption $A(m, q)$ depending on a threshold of non-stationarity θ_{INS} as

$$A(m,q) = \begin{cases} F(q).\frac{L(m)-S}{1+\exp(-k.(d(q)-d_0))} + S, & \delta_{INS} \leq \theta_{INS}; \\ \frac{L'}{1+\exp(-k'.(d(q)-d'_0)}, & \delta_{INS} > \theta_{INS}, \end{cases} \tag{5}$$

where d_0 and d'_0 are the inflection points with corresponding growth rate of k and k'. The S stands for a minimum shift in order to avoid total absorption of signal frames. Moreover $L(m)$ and L' are the maximum absorption values. The $L(m)$ is updated as

$$L(m) = p\delta_{INS} + (1-p)L(m-1), \tag{6}$$

where p assigns the importance of the present RG signal. The second term is defined as the factor $F(q) = d(q)^{1.2-d(q)}$ to guarantee that $A(m, q) \approx L(m)$ only for $d(q) \approx 1$, which indicates an important speech region.

The processed signal $s'(n)$ is obtained by overlap add process of absorbed frames defined as $s'_{frm}(q, n) = A(m, q).\gamma.\tilde{y}_q(n)$.

III. Experiments and Results

In this Section, the proposed NNESE+ARA technique and baseline approaches NNESE [3], ARA_{NSD} [8], SCOE [7] and RSE [9] are evaluated in terms of intelligibility and quality considering several noisy-reverberant conditions. A subset of 200 signals from the IEEE sentences [16] are randomly selected to compose each scenario, which leads to a total of 1200 tests per method. The database consists of male recordings and is chosen for its phonetic balanced sentences in English. Each speech segment is sampled at 16 kHz and has, on average, 2.6 seconds. The room LASP1 is selected from LASP_RIR and the rooms Meeting and Stairway from AIR database. This rooms are selected to represent real urban environments. Rooms Meeting, LASP1 and Stairway presents T_{60} and DRR values of $\{0.36, 0.65, 1.00\}$ and $\{2.7, -3.1, -3.4\}$, respectively. The Meeting room has the smallest T_{60} and

TABLE I

AVERAGE ASII$_{ST}$ INTELLIGIBILITY SCORE [%] FOR ROOMS MEETING, LASP1 AND STAIRWAY WITH NOISES BABBLE AND CAFETERIA.

		Meeting ($T_{60} = 0.36$ s)				LASP1 ($T_{60} = 0.65$ s)				Stairway ($T_{60} = 1.0$ s)			
SNR (dB)		-2	0	2	**Avg.**	-2	0	2	**Avg.**	-2	0	2	**Avg.**
Babble	UNP	45.1	51.1	57.3	51.2	45.7	51.7	58.0	51.8	30.3	35.0	40.1	35.1
	NNESE	60.0	64.9	70.0	65.0	58.2	62.6	66.7	62.5	38.3	41.6	44.8	41.5
	ARA$_{\text{NSD}}$	72.7	75.7	78.2	75.5	68.4	70.0	72.1	70.2	44.0	45.4	46.8	45.4
	SCOE	60.3	67.5	74.7	67.5	58.3	63.7	69.8	63.9	38.1	42.5	46.5	42.4
	RSE	72.3	74.7	76.9	74.6	66.9	68.5	69.8	68.4	35.2	36.2	37.4	36.2
	NNESE+ARA	**77.0**	**79.8**	**81.8**	**79.5**	**69.8**	**71.0**	**72.7**	**71.2**	**44.4**	**46.0**	**46.9**	**45.8**
Cafeteria	UNP	47.9	54.1	60.6	54.2	48.3	54.4	61.0	54.6	32.2	37.0	42.3	37.2
	NNESE	63.0	67.9	72.7	67.9	60.8	64.8	68.8	64.8	40.4	43.6	46.5	43.5
	ARA$_{\text{NSD}}$	73.8	76.8	79.4	76.6	68.6	70.8	72.9	70.7	45.0	46.4	47.7	46.4
	SCOE	65.1	71.8	79.1	72.0	62.9	68.6	74.0	68.5	41.7	46.3	49.3	45.8
	RSE	75.1	77.5	79.7	77.4	67.8	69.2	70.4	69.1	37.2	38.6	39.6	38.5
	NNESE+ARA	**79.8**	**81.2**	**82.8**	**81.3**	**71.1**	**72.8**	**74.0**	**72.6**	**46.3**	**47.4**	**49.4**	**47.7**

highest DRR values. On the other hand, the Stairway is the most challenging condition with the highest T_{60} and lowest DRR. The Babble and Cafeteria additive background noises are selected, respectively, from the RSG-10 [17] and DEMAND [18] databases. Both noises are characterized with non-stationary behavior obtaining maximum INS values of 39 and 23 for signal duration of three seconds [3], respectively.

Speech signals are corrupted considering three SNRs: -2 dB, 0 dB and 2 dB. These values are measured for the reverberated unprocessed speech and the background noise. Intelligibility measures are normalized by the scores achieved for the clean unprocessed signal corrupted by speech shaped noise at 10 dB, considered here as a good intelligibility reference. All scenarios are developed to ensure ASII$_{\text{ST}}$ lowest and highest values between 45.0 and 75.0 for the unprocessed (UNP) speech signal. These scores can be considered as thresholds of poor and good intelligibility [19][20]. The ASII$_{\text{ST}}$ [11] and ESII [12] measures are adopted for the intelligibility evaluation under non-stationary noisy-reverberant conditions. The direct path speech signal, characterized by the first impulse present on each RIR is chosen as the reference signal.

All techniques are applied on a 32 ms frame-by-frame basis. NNESE noise estimation parameters are set to $\alpha = 0.35$ and $\beta = 0.65$. The ARA$_{\text{NSD}}$ operates with a threshold of non-stationarity $\theta_{INS} = 0.4$ and the RG importance $p = 0.7$. Its maximum value for relevant speech regions L' is set to 1.2 and sigmoid parameters are fixed to $k = 17$ for $d = -0.2$ and $k' = 13$ for $d' = 0.5$ with a minimum shift of $S = 0.05$. The SCOE method is performed with four HMM states and the Wiener gain spectral subtraction as in [7]. RSE inverse filtering dereverberation is set to 250 interactions and its spectral subtraction scaling factor to 0.05. Besides the additional normalization step, the NNESE+ARA also performs adopting different sigmoid functions with corresponding parameters of $L' = 1.3$, $k = 17$ for $d = -0.4$ and $k' = 15$ for $d' = -0.3$.

The ASII$_{\text{ST}}$ scores are presented in Table I. Each column corresponds to a room, ordered by the ascending value of T_{60}. Lines are organized for each noise case and corresponding processed method. The NNESE+ARA obtains the best ASII$_{\text{ST}}$ values for all SNR conditions, followed by ARA$_{\text{NSD}}$ and RSE in most of the cases. For the most non-stationary Babble noise, the NNESE+ARA approach achieves the highest ΔASII$_{\text{ST}}$ intelligibility improvements for the Meeting, LASP1 and Stairway rooms with an average gain of 28.3, 20.6 and 10.7, respectively. The RSE and ARA$_{\text{NSD}}$ techniques attain similar results for rooms Meeting and LASP1 with overall averages of 71.5 and 72.8, which indicate mean gains of 20.0 and 21.3. However, as elucidated in [9], the RSE contribution on higher T_{60} are not as expressive due to the length of the inverse filter, which is too short to cover long room impulse responses.

In the Cafeteria scenario, the proposed NNESE+ARA technique also accomplishes the best ASII$_{\text{ST}}$ scores in all SNRs for all rooms. The highest intelligibility gain of 31.9 over all conditions is observed for the Meeting room at -2 dB with ASII$_{\text{ST}}$ values varying from 47.9 up to 79.8. On average, the proposed method attains a intelligibility score of 67.8 over all conditions for the Cafeteria noise compared to 64.6 from the ARA$_{\text{NSD}}$ followed by 62.1 from the SCOE. The NNESE speech enhancement technique applied alone is also able to improve speech intelligibility under noisy-reverberant conditions, attaining an average gain of 13.8, 10.7 and 6.4 for Meeting, LASP1 and Stairway rooms, respectively.

The ESII intelligibility values for each room and noise pair condition is presented in Fig. 2. The proposed NNESE+ARA achieves the highest ESII scores for most challenging conditions of low SNR. Considering the Meeting room, the NNESE+ARA presents the highest average improvement of 0.29 and 0.28 for Babble (a) and Cafeteria (b) noises, respectively. In this room, the average over all ESII improvements for NNESE, ARA$_{\text{NSD}}$, SCOE and RSE are 0.12, 0.23, 0.17 and 0.23. For the LASP1 room, the proposed method obtained on average a ΔESII gain of 0.19 and 0.18, compared to 0.18 and 0.16 for the ARA$_{\text{NSD}}$ approach followed by 0.16 and 0.14 for the RSE technique. The highest T_{60} present on the Stairway room leads to a more challenging condition. In this scenario, the highest intelligibility score of 0.36 is achieved by the SCOE for both noises at 2 dB. However, considering all SNR values, the NNESE+ARA and ARA$_{\text{NSD}}$ accomplish ESII average values of 0.34 and 0.33 compared to 0.32 for

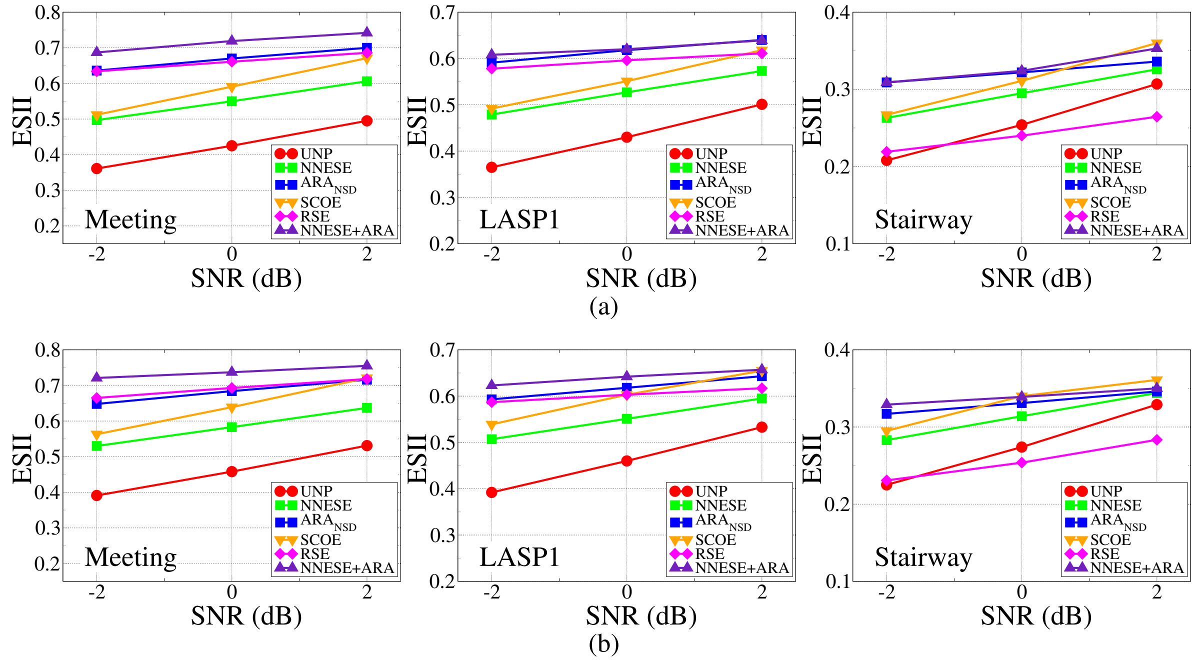


Fig. 2. Average ESII intelligibility scores for rooms Meeting, LASP1 and Stairway considering acoustic noises Babble (a) and Cafeteria (b).

TABLE II

PESQ SCORES FOR MEETING ROOM.

	Babble				Cafeteria			
SNR (dB)	-2	0	2	Avg.	-2	0	2	Avg.
UNP	2.06	2.08	2.22	2.12	2.21	2.33	2.47	2.34
NNESE	2.07	2.14	2.25	2.15	2.24	2.37	2.50	2.37
ARA_{NSD}	2.19	2.30	2.16	2.22	2.30	2.33	2.42	2.35
SCOE	2.18	2.27	2.43	2.29	2.31	2.45	2.61	2.46
RSE	2.12	2.17	2.27	2.19	2.23	2.30	2.44	2.32
NNESE+ARA	**2.36**	**2.33**	**2.47**	**2.39**	**2.56**	**2.61**	**2.69**	**2.62**

TABLE III

PESQ SCORES FOR LASP1 ROOM.

	Babble				Cafeteria			
SNR (dB)	-2	0	2	Avg.	-2	0	2	Avg.
UNP	2.01	2.03	2.10	2.05	2.04	2.14	2.31	2.16
NNESE	2.07	2.11	2.15	2.11	2.13	2.21	2.39	2.24
ARA_{NSD}	2.16	2.18	2.24	2.19	2.26	2.24	2.35	2.28
SCOE	2.15	2.22	2.28	2.22	2.19	2.28	2.47	2.31
RSE	2.01	2.08	2.21	2.10	2.01	2.10	2.26	2.12
NNESE+ARA	**2.27**	**2.32**	**2.50**	**2.36**	**2.56**	**2.62**	**2.59**	**2.59**

TABLE IV

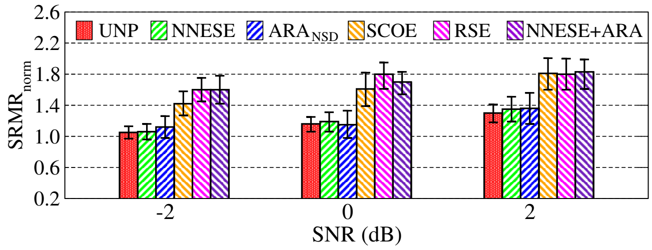
PESQ SCORES FOR STAIRWAY ROOM.

	Babble				Cafeteria			
SNR (dB)	-2	0	2	Avg.	-2	0	2	Avg.
UNP	1.75	1.85	2.09	1.90	1.83	2.07	2.01	1.97
NNESE	1.92	2.02	2.08	2.01	2.05	2.32	2.15	2.17
ARA_{NSD}	1.93	2.12	2.16	2.07	1.98	2.37	2.20	2.19
SCOE	1.98	2.14	2.31	2.14	2.03	2.28	2.35	2.22
RSE	1.94	1.99	2.13	2.02	1.97	2.08	2.05	2.03
NNESE+ARA	**2.27**	**2.33**	**2.35**	**2.32**	**2.58**	**2.56**	**2.56**	**2.56**

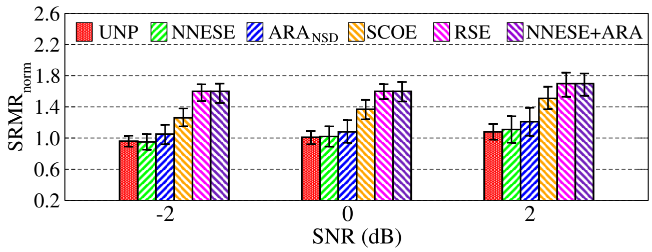
the SCOE method. These values correspond to a 26%, 24% and 23% increments on intelligibility.

The objective quality assessment for each method under noisy-reverberant conditions is performed based on the PESQ [21] and $SRMR_{norm}$ [13][22]. The PESQ is computed considering 60 frames uniformly distributed over the symmetrical distance and the $SRMR_{norm}$ adopts 256 ms rate with 87.5% overlap. Table II show the PESQ values acquired by each method for the Meeting room. The NNESE+ARA attains the highest scores for all SNR cases. In the Babble scenario of the Meeting room, the technique achieves an average PESQ of 2.39, followed by 2.29 and 2.22 from SCOE and ARA_{NSD}. Considering the Cafeteria noise, NNESE+ARA also accomplishes the highest scores of 2.56, 2.61 and 2.69 for SNRs of -2 dB, 0 dB and 2 dB, accordingly. In this context, the SCOE presents values of 2.31, 2.45 and 2.61 followed by the NNESE scores of 2.24, 2.37 and 2.50. The LASP1 room results are presented in Table III. The best performance is achieved by NNESE+ARA, SCOE and ARA_{NSD}. These approaches attain an average result of 2.36, 2.22 and 2.19 for the Babble noise and 2.59, 2.31 and 2.28 for the Cafeteria. The RSE presents PESQ gain for small SNR values, which is justified by the fact that it does not explicitly take into account acoustic noises on its spectral suppression step. In Table IV the Stairway room results are presented. The NNESE+ARA achieves the best average PESQ results of 2.32 and 2.56 for Babble and Cafeteria noise, respectively. Furthermore, the SCOE method presents equivalent values of 2.14 and 2.22, followed by the ARA_{NSD} technique with 2.07 and 2.19. The corresponding values for the NNESE technique alone are 2.01 and 2.17 in this context. These results reinforce the capacity of NNESE and ARA_{NSD} to jointly deal with noise-reverberant distortions. Moreover, results demonstrate that NNESE+ARA, SCOE and ARA_{NSD} are generally superior in the PESQ quality sense considering the scenarios adopted for analysis.

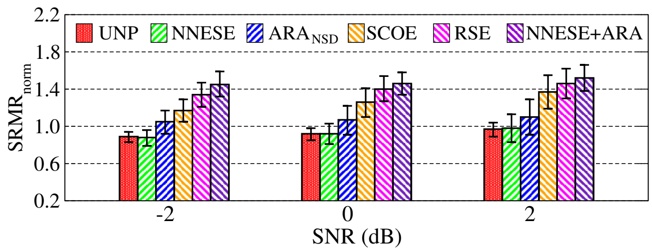
Figure 3 illustrates the average $SRMR_{norm}$ values over the two noises for rooms Meeting, LASP1 and Stairway. The goal is to distinguish among the five approaches the ones that can better mitigate temporal coloration on speech signals. The NNESE+ARA and RSE methods present the highest quality $SRMR_{norm}$ scores for most of the scenarios. This is the case for the Meeting room (a) at -2 dB, in which both techniques attain $SRMR_{norm} = 1.60$. Slightly better results are achieved by the RSE at 0 dB and by NNESE+ARA at 2 dB, the two



(a)



(b)



(c)

Fig. 3. Average SRMR$_{norm}$ over noises for rooms Meeting (a), LASP1 (b) and Stairway (c).

examples demonstrate the highest quality value of 1.80 and 1.81, respectively. In the LASP1 scenario (b), these approaches present similar behavior with the SRMR$_{norm}$ scores of 1.62, 1.62 and 1.71 for -2 dB, 0 dB and 2 dB, respectively. These values equal quality increments of 64%, 59% and 54%. For the Stairway room (c), the NNESE+ARA accomplishes the best quality results in all scenarios. The proposed method attains SRMR$_{norm}$ scores of 1.45, 1.46 and 1.52 compared to 1.34, 1.40 and 1.46 for the RSE approach followed by values of 1.17, 1.26 and 1.37 obtained by the SCOE technique. These results reinforce the capacity of the proposed method to provide intelligibility and quality improvement in noisy-reverberant environments.

IV. Conclusion

In this paper, a two-stage time domain approach was introduced to improve intelligibility of speech signals under noisy-reverberant conditions. The NNESE and ARA$_{NSD}$ techniques were adapted and jointly taken into account to mitigate the effects of noise and reverberation separately. The NNESE+ARA obtained the highest ASII$_{ST}$ results for all noisy-reverberant conditions considering two non-stationary acoustic noises and three rooms. A similar behavior was observed for the ESII objective measure in most cases. It was shown that the NNESE+ARA also attains quality improvements achieving best PESQ and average SRMR$_{norm}$ results for most of the reverberant rooms considered in the experiments.

References

[1] R. Bolt and A. MacDonald, "Theory of speech masking by reverberation," *The Journal of the Acoustical Society of America*, vol. 21, no. 6, pp. 577–580, 1949.

[2] A. Nabelek, "Communication in noisy and reverberant environments," *Acoustical factors affecting hearing aid performance*, pp. 15–28, 1993.

[3] R. Tavares and R. Coelho, "Speech enhancement with nonstationary acoustic noise detection in time domain," *IEEE Signal Processing Letters*, vol. 23, no. 1, pp. 6–10, 2016.

[4] L. Zão, R. Coelho, and P. Flandrin, "Speech enhancement with emd and hurst-based mode selection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 5, pp. 899–911, 2014.

[5] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[6] R. Coelho and L. Zão, "Empirical mode decomposition theory applied to speech enhancement," in *Signals and Images: Advances and Results in Speech, Estimation, Compression, Recognition, Filtering and Processing* (R. Coelho, V. Nascimento, R. Queiroz, J. Romano, and C. Cavalcante, eds.), Boca Raton, Florida: CRC Press, 2015.

[7] C. S. J. Doire, M. Brookes, P. A. Naylor, C. M. Hicks, D. Betts, M. A. Dmour, and S. H. Jensen, "Single-channel online enhancement of speech corrupted by reverberation and noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 572–587, March 2017.

[8] G. Zucatelli and R. Coelho, "Adaptive reverberation absorption using non-stationary masking components detection for intelligibility improvement," *IEEE Signal Processing Letters*, vol. 27, pp. 1–5, 2020.

[9] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 3, 2006.

[10] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *16th International Conference on Digital Signal Processing*, pp. 1–5, 2009.

[11] R. C. Hendriks, J. B. Crespo, J. Jensen, and C. H. Taal, "Optimal near-end speech intelligibility improvement incorporating additive noise and late reverberation under an approximation of the short-time sii," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 5, pp. 851–862, 2015.

[12] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2181–2192, 2005.

[13] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 55–59, IEEE, 2014.

[14] D. Pastor and F. Socheleau, "Robust estimation of noise standard deviation in presence of signals with unknown distributions and occurrences," *IEEE Transactions on Signal Processing*, vol. 60, pp. 1545–1555, April 2012.

[15] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, and J. Xiao, "Testing stationarity with surrogates: A time-frequency approach," *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3459–3470, 2010.

[16] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[17] H. J. Steeneken and F. W. Geurtsen, "Description of the rsg-10 noise database," *report IZF*, vol. 3, p. 1988, 1988.

[18] J. Thiemann, N. Ito, and E. Vincent, "Demand: a collection of multi-channel recordings of acoustic noise in diverse environments," in *Proc. Meetings Acoust.*, 2013.

[19] "American National Standard Methods for Calculation of theSpeech Intelligibility Index," standard, American National Standard Institute, Mar. 1997.

[20] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, pp. I–I, IEEE, 2006.

[21] I. 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.

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