ICASSP2021 TORONTO Canada . June 6-11, 2021

Abstract

This letter introduces a new time domain absorption approach designed to reduce masking components of speech signals under noisyreverberant conditions. In this method, the non-stationarity of corrupted signal segments is used to detect masking distortions based on a defined threshold. The non-stationarity is objectively measured and is also adopted to determine the absorption procedure. Additionally, no prior knowledge of speech statistics or room information is required for this technique. The results show that the proposed scheme leads to a higher intelligibility improvement when compared to competing methods for three objective measures. A perceptual listening test is also considered and corroborates these results.

Reverberation and Non-Stationarity

The natural non-stationary behavior of speech signals is closely related to speech intelligibility. The main goal of adaptive methods is to preserve transient speech regions in order to improve speech intelligiblity. The reverberation effect is usually described by a Room Impulse Response h(n). In real environments, acoustic noises w(n) are also a common distortion, which means that both effects can occur simultaneously. Therefore, the resultant noisy-reverberant speech signal s(n) can be obtained by:



Fig. 1. Spectrogram and related INS for direct signal (top), reverberated • As INS alters on noisy-reverberant scenarios, it can be a useful signal with $T_{co} = 4.9$ s and SRR=7.1 dB (middle), and reverberated signal with instrument for detection and reduction of such effects. Chainsaw noise at -3 dB (bottom)

ARA The Adaptive Reverberation Absorption with Non-Stationary Detection (ARA_{NSD}): The approach works similar to a physical element, changing the low absorption characteristic of materials that compose a room. The proposed technique can be described in two main phases: reverberation detection and acoustic absorption.



Fig. 2. Non-stationarity variation $\delta_{\mu\nu}(m)$ for the reverberated and noisyreverberated signals.

Absorption

Implemented on a frame-by-frame basis and established depending on the value of $\theta_{\rm me}$, the theshould of non-stationarity. For each frame *I*, a INS vector $v_{frm}(l)$ is extracted and a short-time distance d(l) is computed, similarly as in the RG case.



Adaptive Reverberation Absorption using Non-Stationary Masking Components Detection for Intelligibility Improvement

G. Zucatelli and R. Coelho

Laboratory of Acoustic Signal Processing (LASP) Military Institute of Engineering (IME), Rio de Janeiro, Brazil zucatelli@ime.eb.br, coelho@ime.eb.br lasp.ime.eb.br

$$s(t) = h(t) \circledast x(t) + w(t)$$

The Index of Non-Stationarity (INS): In this work, the INS is adopted to objectively examine the non-stationarity of speech signals under noisyreverberant environments.

This measure is a time-frequency approach that compares the target signal with stationarity references called surrogates for different time scales T_h/T , where T_h is the short-time spectral analysis length and T is the total signal duration. For each length T_{h} , a threshold γ is defined to keep the stationarity assumption considering a 95% confidence degree:

$$\left. \begin{array}{ll} \mathsf{INS} \begin{cases} \leq \gamma : & \mathsf{stationary,} \\ > \gamma : & \mathsf{non-stationary} \end{cases} \right. \end{array} \right.$$

Important Notes:

• Reverberation and acoustic noise significantly change the temporal and spectral structure of speech signal.

• Natural non-stationary behavior of speech signal can be considerably attenuated (INS: $200 \rightarrow 100$).

Reverberation Groups (RG): The m-th segment composed of eight consecutive frames of the reverberant speech $s_{PC}(m,n)$.

This window duration is selected to enable a long-term temporal observation of the reverberation effect using the INS.

INS vector (v_{ms}) : For each $s_{RG}(m,n)$, the INS values are grouped into $v_{\mu\nu}$ vectors. which characterizes the non-stationary behavior of the RG.

Consecutive vectors are then used to compute a normalized variation $\delta_{\mu\nu}(m)$ of the non-stationary property as:

$$\delta_{INS}(m) = \frac{||\mathbf{v}_{\mathbf{INS}}(m) - \mathbf{v}_{\mathbf{INS}}(m-1)||}{||\mathbf{v}_{\mathbf{INS}}(m)|| + ||\mathbf{v}_{\mathbf{INS}}(m-1)||}$$

Absorption: Sigmoid functions are selected to assign each value of d(l) to a corresponding absorption A(m,l) because of their smoothness and monotonic property. The proposed adaptive absorption A(m, l) is therefore defined in every frame l by

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

The resulting signal is reconstructed by overlap-add steps of processed frames.

 240 speech signals from TIMIT database • Reverberation database AIR for rooms Stairway (T_{60} =1.1 s) and Aula Carolina (T_{60} =4.9 s) • Four SNR conditions (-3dB, -2dB, -1dB, 0dB, 1dB) for Babble, Cafeteria, Chainsaw and SSN

Non-Stat. Restored

The ARA_{NSD} is able to absorb masking components of the corrupted signal, e.g. near 0.7 s and 1.5 s, which makes the resulting signal more similar to its anechoic version.

Experimental Setup:

Moreover, the proposed method restores the natural nonstationarity behavior raising the INS value from 100 up to 150, which is closer to the direct signal (INS=200).

Objetive Evaluation

The ESII and ASII_{st}: These measures are adopted for the intelligibility prediction because they are explicitly designed to deal with the nonstationarity of speech and its distortions. SSS G-D AGC A-A ARA_{NSD}

								Q
		ESII (%)						
R ((dB)	-3	-2	-1	0	1	Avg.	
0	UNP	63.8	64.8	65.7	66.6	67.3	65.6	
NS=4	SSS	65.7	66.7	67.4	68.3	68.9	67.4	
	AGC	66.9	67.9	68.6	69.3	69.9	68.5	8
	ARA_{NSD}	68.8	69.2	69.5	69.7	69.9	69.4	
_	UNP	65.1	66.0	66.7	67.3	68.0	66.6	
1	SSS	67.2	67.9	68.5	69.0	69.6	68.4	
2 Z	AGC	68.2	68.9	69.6	70.2	70.6	69.5	
<u> </u>	ARA _{NSD}	69.6	69.8	70.1	70.3	70.6	70.1	12
_	UNP	56.9	58.2	59.5	60.6	61.8	59.4	
ĭ	SSS	58.9	60.2	61.5	62.5	63.7	61.3	
ZN N	AGC	60.3	61.6	62.8	64.0	65.1	62.8	
=	ARA _{NSD}	67.2	67.6	68.1	68.4	68.8	68.0	$ 0^{-7} -3 -2$
	UNP	62.2	63.4	64.5	65.4	66.3	64.4	
NS=2	SSS	64.4	65.6	66.6	67.4	68.2	66.4	
	AGC	65.7	66.7	67.7	68.6	69.3	67.6	4
	ARA _{NSD}	68.6	68.9	69.3	69.6	69.9	69.3	
								0

Fig. 5. ESII and ASII_{st} for Aula Carolina room (T60=4.9 s).

Perceptual Evaluation

• Listening test with 10 male Brazilian volunteers • Simulated room with Image Source Method (ISM) - T_{eo} =1.0 s • SSN noise at -3dB, 0dB, 3dB.



This letter proposed a new time domain absorption approach designed to reduce masking components of speech signals under noisyreverberant conditions. In this method, the non-stationarity of corrupted signal segments are assessed via INS and used to detect masking distortions based on a defined threshold. The results showed that the proposed scheme leads to a higher intelligibility improvement when compared to competing methods. A perceptual listening test was also developed and corroborate the objective results.

R. Bolt and A. MacDonald, "Theory of speech masking by reverberation," J. Acoust. Soc. Amer., vol. 21, no. 6, pp. 577–580, 1949.

T. Arai, N. Hodoshima, and K. Yasu, "Using steady-state suppression to improve speech intelligibility in reverberant environments for elderly listeners," IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 7, pp. 1775–1780, Sep. 2010. P. N. Petkov and Y. Stylianou, "Adaptive gain control for enhanced speech intelligibility under reverberation," IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1434–1438, Oct. 2016. P. Borgnat, P. Flandrin, P. Honeine, C. Richard, and J. Xiao, "Testing stationarity with surrogates: A time-frequency approach," IEEE Trans. Signal Process., vol. 58, no. 7, pp. 3459–3470, Jul. 2010. L. Zão, R. Coelho, and P. Flandrin, "Speech enhancement with EMD and hurst-based mode selection," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 22, no. 5, pp. 899–911, May 2014.

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, FL, USA: CRC Press, 2015. R. Tavares and R. Coelho, "Speech enhancement with nonstationary acoustic noise detection in time domain," IEEE Signal Process. Lett., vol. 23, no. 1, pp. 6–10, Jan. 2016.

□ S. Ghimire, "Speech intelligibility measurement on the basis of ITU-T Recommendation P.863," 2012.











