Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Wiener Gain and Deep Neural Networks: A Well-Balanced Pair For Speech Enhancement

Dayana Ribas (dribas@unizar.es)
University of Zaragoza
Antonio Miguel
University of Zaragoza
Alfonso Ortega
University of Zaragoza
Eduardo Lleida
University of Zaragoza

Research Article

Keywords: Wiener gain estimator, Speech enhancement, Noise reduction, DNN, OMLSA

Posted Date: September 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-900751/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Wiener Gain and Deep Neural Networks: A well-balanced pair for speech enhancement.

Dayana Ribas*, *AntonioMiguel*, *AlfonsoOrtega*, *EduardoLleida*
dribas@unizar.es

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain

Abstract

This paper proposes a Deep Neural Network (DNN)-based Wiener gain estimator for speech enhancement. The proposal is in the framework of the classical spectral-domain speech enhancement algorithms. In this case, we used the Optimal Modified Log-Spectral Amplitude (OMLSA), but consider that this proposal could fit many alternative speech estimation algorithms. We determined the best usage of the DNN approach at learning a robust instance of the Wiener gain estimator according to the characteristics of the SNR estimation and the gain function. To design a DNN architecture adjusted for the speech enhancement task, we study various configuration issues frequently used in DNN-based solutions, including speech representations, residual connections, and causal vs. non-causal designs. Thus, we provide conclusions for the use of DNN architectures with the enhancement purpose. Experiments show that the proposal provides results on the state-of-the-art. But beyond the objective quality measures, there are examples of noisy vs. enhanced speech available for listening to demonstrate in practice the skills of the method in real audio.

Keywords: Wiener gain estimator Speech enhancement Noise reduction DNN OMLSA

1. Introduction

Speech enhancement (SE) has been the target of many research efforts for several decades. Advances in the understanding of environmental acoustic distortion on speech signals during this period have motivated the development of many SE methods [1]. The scale of the problem differs whether signals are single or multichannel. The availability of multiple channels provides an acoustic reference that could be used on the enhancement processing, while the single-channel case represents the major challenge. However, there are applications, for instance, telephonic signals, that have no choice but to perform in this condition.

Single-channel SE has been traditionally performed using statistical methods. An established approach is the family of spectral-domain speech enhancement algorithms, which are based on the gain-based approach [1, 2]. In this framework, the noisy speech is transformed into the time-frequency (t-f) domain. Then an estimated gain is applied to the t-f speech representation in order to obtain an enhanced version. Finally, a synthesis stage applies an inverse transformation to the enhanced t-f speech to obtain the signal back into the time domain. Some of best-known methods among this framework are the Wiener filter [3], the Spectral Subtraction (SS) [4], the Short-Time Spectral Amplitude (STSA) [5], and the Log-Spectral Amplitude estimator (LSA) [6]. In general, these approaches rely on estimations of the *a priori* Signal-to-Noise Ratio (SNR), which are used to compute the gain function to determine the attenuation of noise-dominated t-f regions. Although many SE algorithms follow the gain-based approach, they mainly differ in the way the *a priori* SNR is estimated and in which gain function they use. The design of the gain function and the accuracy of the *a priori* SNR estimation can become the main weakness of the SE method. In realistic scenarios, the dynamic fast changes of non-stationary impulsive noise and the mixture of noise types including speech-correlated noises, propose significant challenges for statistical SNR estimators [7].

The availability of deep speech enhancement solutions has grown fast during the last few years. The paradigm of data-driven methods might be a suitable solution to handle the complex process of acoustic speech distortion. Since 2012, many solutions for SE based on Deep Neural Network (DNN) architectures have been

proposed. These algorithms follow different strategies that usually lie on top of the gain-based approach that also rules the traditional SE statistical methods [1], but adapted to the deep learning routine.

There are previous works studying different DNN-based estimations of the SNR and the gain function. The work of Xia et al. [8, 9] was the first approach to the DNN-based SE by supporting the Wiener filtering with a Weighted Denoising Autoencoder, that estimates the clean speech by subband and then uses it for estimating the short-term *a priori* SNR and the filter gain function. Similarly, from the mono-aural speech separation the t-f masking approach with Ideal Binary/Ratio Masking (IBM/IRM) has been used for performing feature enhancement in Automatic Speech Recognition (ASR) [10, 11]. In [12, 13] authors propose a supervised learning algorithm for IRM estimation to perform noise-robust ASR. Then, [14, 15, 16, 17] extensively used the DNN-based estimation of IBM and IRM for hearing aids purposes also applied to ASR. For cochlear implants applications, [18, 19, 20] extended the SE using DNN based on IRM to novel speakers and proposed an approach suitable for practical applications with low latency.

In this paper, we develop a gain-based approach for SE that provides robustness using deep learning. We propose a DNN-based architecture to estimate the Wiener gain function and perform SE supported on the classical LSA speech estimation, specifically in its Optimally Modified version (OMLSA) [21]. First, we studied the key points of the gain-based SE algorithm and accordingly, the best use of a deep learning solution. We found that DNNs can provide a more robust estimation at learning the Wiener gain estimator for previously unseen noisy conditions than other approaches. This fact highlights the main novelty of the paper with respect to the mentioned previous work, that uses the DNN for different purposes in the algorithm. Furthermore, we study various configuration issues in the DNN architecture used for SE. We addressed the speech representations, the use of residual connections, and online vs. offline designs for practical applications.

This work first contributes with a data-driven Wiener gain estimator for SE that can be generalized to different approaches of the classical spectral-domain speech enhancement algorithm. It also demonstrates the usefulness of deep learning for expanding the application scope of established SE schemes in realistic scenarios with challenging environmental noises. Further, this paper aims to contribute with conclusive guidelines of the suitability and practical usability of DNN-based SE solutions. Examples of enhanced signals are available [22].

In the following, section 2 introduces the speech enhancement task through the spectral-domain speech enhancement algorithm. Then, section 3 describes the proposal followed by the experimental setup in section 4. Section 5 presents an analysis of some design issue on the speech enhancement based on a DNN architecture. Finally, section 6 presents results and discussion, and section 7 concludes the paper.

2. Speech enhancement

Let $y(n)$ denote the observed noisy speech signal given by $y(n) = x(n) + d(n)$ with $x(n)$ the clean speech, $d(n)$ the additive noise, and n the discrete-time index. The pre-processing stage for performing speech enhancement in the spectral domain starts with a short-term speech analysis of the segmented $y(n)$ into overlapping frames through the application of a window function. Then, a short-term Fourier transform (STFT) is used to obtain the spectral representation:

$$Y(k,l) = \sum_{n=0}^{N-1} y(n+lM)h(n)e^{-j(2\pi/N)nk}, \qquad (1)$$

where l is the time frame index and k the frequency bin index, $h(n)$ is the analysis window of size N, and M is the window shift.

Figure 1 depicts the spectral-domain speech enhancement algorithm. The power spectrum $|Y(k,l)|^2$ is used as input of the noise reduction block, while the spectral phase is kept apart for the last block of post-processing for speech reconstruction.

The core of the enhancement method is depicted in the central block. This approach is the paradigm followed by the family of spectral-domain speech enhancement algorithms. However, notice that in figure 1 we present the case where the hypotheses of speech presence and absence are separately considered [23, 21]. The spectral-domain speech enhancement algorithm uses $|Y(k,l)|^2$ for computing the spectral Gain, which is used to obtain the filter that modifies $|Y(k,l)|^2$ according to the speech presence probability (See section II in [5]).

2.1. Speech estimation algorithms

Many speech estimation algorithms follow the aforementioned scheme, for instance, the well-known approaches Wiener filter, SS [4], STSA [5], and Minimum Mean-Square Error Log-Spectral Amplitude estimator (LSA) and its modified versions [6, 24, 21]. The main difference among these speech estimation methods is

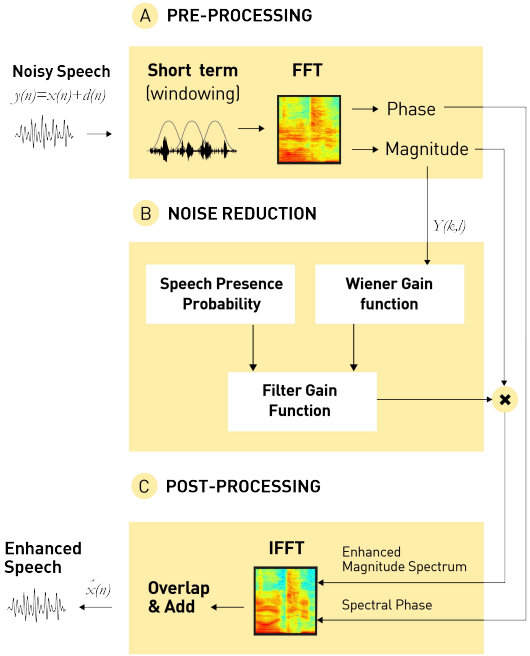


Figure 1: *spectral-domain speech enhancement algorithm*

the filter gain function. In the case of the Wiener filter, which is the MMSE estimator of the clean speech subject to a linear constraint, the spectral gain is:

$$G_W(k,l) = \frac{\xi_{k,l}}{1+\xi_{k,l}}, \tag{2}$$

where $\xi_{k,l}$ is the *a priori* SNR computed with the clean and noisy speech for each frequency bin k and each time segment l.

From this statement, we can define other speech estimation algorithms in terms of the $G_W(k,l)$. For instance, in the SS algorithm the gain function is the square root of the maximum likelihood estimator of each spectral component variance [23]. In terms of $G_W(k,l)$ it can be expressed as:

$$G_{SS}(k,l) = \sqrt[\beta]{G_W(k,l)} \tag{3}$$

with $\beta = 2$. Albeit, several modifications of this algorithm have been studied in terms of changing the value of β [1].

For the LSA family the gain function also depends on $G_W(k,l)$ [6]

$$G_{LSA}(k,l) = G_W(k,l)\exp(\frac{1}{2}\int_{\upsilon(k,l)}^{\infty}\frac{e^{-t}}{t}dt) \tag{4}$$

where

$$\upsilon(k,l) = \frac{|Y(k,l)|^2}{\lambda_d(k,l)}G_W(k,l) \tag{5}$$

with $\lambda_d(k,l) = E\{|D(k,l)^2|\}$ the variance of the kth spectral component of the noise for frame l. Thus, different modifications of the LSA estimator also express the gain function in terms of $G_W(k,l)$. For instance the Gain function of the optimally modified version of LSA is

$$G_{OMLSA}(k,l) = G_{LSA}^{p(k,l)}(k,l)G_{min}^{1-p(k,l)} \tag{6}$$

where G_{min} is a gain lower bound threshold, and $p(k,l)$ is the speech presence probability. This optimally modified gain function outperforms previous alternatives of the spectral-domain speech enhancement [21].

As we can see, $G_W(k,l)$ is a common element of utmost importance in the gain function of classical speech estimation algorithms. However, its dependence on the *a priori* SNR (ξ_k) makes it sensitive to errors, because it is not directly accessible from the observed spectral power $|Y(k,l)|^2$ and has to be estimated for each segment l.

2.2. SNR Estimation

Classical speech enhancement algorithms are commonly described in terms of the *a priori* and *a posteriori* SNR [5, 6, 21, 1]. The *a priori* SNR is defined in terms of the Power Spectral Density (PSD) of the clean speech and the noise signal:

$$\xi_{k,l} = \frac{P_x(k,l)}{P_d(k,l)} \tag{7}$$

where $P_x(k,l) = |X(k,l)|^2$ is the clean speech PSD, $P_d(k,l) = |D(k,l)|^2$ is the noise signal PSD, both in frequency bin k. The *a posteriori* SNR depends on the noise signal PSD and the noisy spectral power $P_y(k,l) = |Y(k,l)|^2$:

$$\gamma_{k,l} = \frac{|Y(k,l)|^2}{P_d(k,l)} \tag{8}$$

As we can see, with an estimate of the PSD of the noise, the *a posteriori* SNR can be directly obtained using the noisy PSD $|Y(k,l)|^2$. Many statistical algorithms have been proposed to estimate the noise spectrum. For instance there are histogram-based approaches [25], minimum statistics [26], Minima Controlled Recursive Averaging (MCRA) [21], etc. However, in general they lose accuracy handling realistic non-stationary noises. Also, they could distort the speech signal or generate annoying artifacts. Besides, the *a priori* SNR also needs an estimate of the PSD of the clean signal, which is another challenging point of these approaches.

3. Proposal

Nowadays, we can take advantage of the data-driven paradigm behind deep learning for modeling the relationship between noise and clean data. A DNN regression can be used in speech estimation algorithms for instance to estimate the instantaneous SNR, for obtaining the *a priori* SNR. However, when the observed signal is barely noise affected, i.e. it is mostly clean, the SNR can rise to ∞. This case, the DNN regression would be more sensitive to errors because there will be a huge amount of possible values to provide as result. However, as the G_W depends on the SNR (equation 2), high SNR conditions provoke high gain values $G_W \rightarrow 1$, while low SNR conditions dump $G_W \rightarrow 0$.

In this paper, we propose the estimation of the Wiener gain (equation 2) by using a DNN. This way, the dynamic range for the DNN regression to obtain G_W would be bounded [0, 1], which is a more accurately achievable task for a DNN.

Furthermore, the use of a DNN in this task was also motivated by the fact that we can implement a causal enhancement system. This means that as this is not necessarily depending on future time frames it can be employed in online applications. Also, the DNN performs non-recursive estimations, which avoid the re-insertion of estimation errors from previous frames. Mentioned previous statistical SNR-estimators are usually based on recursive and non-causal schemes [9, 13].

3.1. Noise reduction algorithm

Figure 2 depicts the DNN-based noise reduction method proposed. The DNN is trained supervisedly with clean speech data and noise patterns through a standard data augmentation process. Thus, inputs are $P_x(k,l)$ and $P_d(k,l)$ and the output is $G_W(k,l)$.

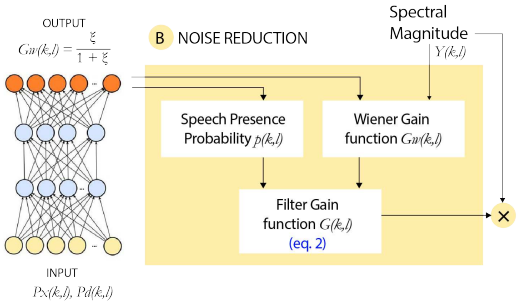


Figure 2: *Noise reduction block based on DNN.*

For computing the final speech enhancement filter gain function $G(k,l)$, the term delivered by the DNN $G_W(k,l)$ is directly used as Wiener Gain function $G_W(k,l)$ and applied according to the speech presence probability $p(k,l)$. Therefore, we will also use the DNN output to feed the speech presence probability block.

4. Experimental setup

4.1. Datasets

Data for DNN training: . We used 16 kHz sampled data from Timit and Librispeech datasets in English, as well as some data in Spanish from Albayzin, Speechdatcar, Tcstar, Mavir, and some hours of Spanish TV emissions. Approximately 120 hours of clean speech. This data was augmented by adding randomly stationary and non-stationary noises from the Musan dataset [27], SNR = 0-30 dB, including music and speech, and scaling the time axis at the feature level. In the following you can find the links to the aforementioned datasets:

- TIMIT: https://catalog.ldc.upenn.edu/LDC93S1
- Librispeech: http://www.openslr.org/12/
- Albayzin: http://catalogue.elra.info/en-us/repository/browse/ELRA-S0089/
- Speechdatcar: http://catalogue.elra.info/en-us/repository/browse/ELRA-S0140/
- : Tcstar: http://www.elra.info/en/projects/archived-projects/tc-star/
- : Mavir: http://www.lllf.uam.es/ESP/CorpusMavir.html

Data for speech enhancement:. We created a simulated noisy dataset with 20958 speech utterances. Using clean read phrases in Spanish, phonetically balanced, from the laboratory sessions of the AV@CAR dataset [28]. The dataset is sampled at 16 kHz and includes 20 speakers considering males and females. The clean data was corrupted with a representative selection of stationary and non-stationary additive noise that could be in everyday scenarios including:

- Babble: Noisy pattern from the talking of many people. It is a special case of non-stationary noise, very difficult to handle because it is highly correlated with the target voice since it is also voice.
- Traffic: Noise from the traffic at a random street, including cars, klaxon, street noise, etc.
- Cafe: Mixture of environmental noises in a cafe, including people talking, noise from cutlery, etc.

- Tram: Environmental noise in a tram station, including some stationary segments when the tram arrives.

Each noisy subset at $SNR = 0, 5, 10, 15, 20$ dB.

4.2. Speech enhancement configuration

To obtain the enhanced speech, the OMLSA method was implemented and parameters were selected according to table 1 in [21]. Waveform reconstruction was performed using the original spectral phase of the corresponding noisy signal.

5. DNN Architecture: Analysis of design issues for speech enhancement

From now on, we focus on defining the DNN architecture for estimating the Wiener gain function $G_W(k, l)$. However, we would like to analyze some issues at designing a DNN architecture with SE purpose. So, we start from a basic Convolutional Neural Networks (CNN) topology with 1-dimensional convolutions (Conv1D) and kernel size $k = 3$, built using the Pytorch toolkit. We use AdamW for training the network [29, 30], PReLUs [31] as parametric nonlinearity and Mean Square Error (MSE) as cost function.

In the following, section 5.1 addresses the selection of speech representations to input the DNN. Section 5.3 discusses the advantage of using residual connections, and section 5.2 approaches the topic of online/offline processing. Finally, we conclude on a DNN architecture design for estimating G_W and proceed to test the performance of the proposed SE method.

5.1. Speech representations

DNN architectures can easily incorporate different speech representations by stacking feature sets in the input network vector. The use of different sources of information provides an advantage to the network processing since each representation can provide an alternative view of the speech signal that enriches the characterization of the corrupted signal structure. This strategy has been employed by many authors in their SE proposals, demonstrating its benefit alongside different application contexts. Accompanying basic spectral transformations, such as Fourier Transform, it is common to add perceptual speech representations, such as Mel-scaled filterbank (FB), Mel Frequency Cepstral Coefficients (MFCC), among others [13, 32].

In order to test the impact of using complementary speech representations, we define a front-end with different features computed on a 25 ms Hamming window frame (shift = 10 ms). For each frame segment, three types of acoustic feature vectors are computed and stacked, to create a single input feature vector for the network: 512-dimensional Fast Fourier Transform (FFT), 32 FB, and 32 MFCC (This feature size was selected considering our experience in previous works [32, 33]). Finally, input data are normalized using the mean and variance of the training dataset. Input features for training the network were generated on-the-fly, operating in contiguous vector blocks of 200 samples, so that convolutions in the time axis can be performed.

Figure 3 presents speech quality results expressed in terms of SNR.

- SNR (dB): Is the initial SNR fixed at mixing speech and noise .
- SNR output (dB): Corresponds to the SNR after the enhancement process. As we are working with simulated speech signals, what we do is to filter apart the clean and noise signals with the enhancement filter ($G_W(k, l)$) for computing the classical SNR. This way, we are able to measure the attenuation of the enhancement filter, by means of the exact SNR value when the speech is enhanced.

We used FFT and three combinations with complementary feature representations: $FFT-FB$, $FFT-MFCC$, $FFT-FB-MFCC$. See that all complementary representations outperform the single use of FFT, providing the SE method with improved results in terms of SNR level. $FFT-FB-MFCC$ reached the best performance among all.

5.2. Online processing: Causality

From the practical application view, the causality is an important topic. Causal designs allow the system to perform an online processing, since it just relies on the present and past observations to compute the enhancement. Conversely, a non-causal design assumes the full speech signal is available, so it employs the previous and posterior frames of a sample to compute the enhancement. The non-causal designs are more frequently used in previous works. However, the assumption that future frames are not available for the computation imposes a limitation for the use in realistic online applications. This section presents a comparison between causal and non-causal design in the framework of the CNN architecture. The difference consists of the

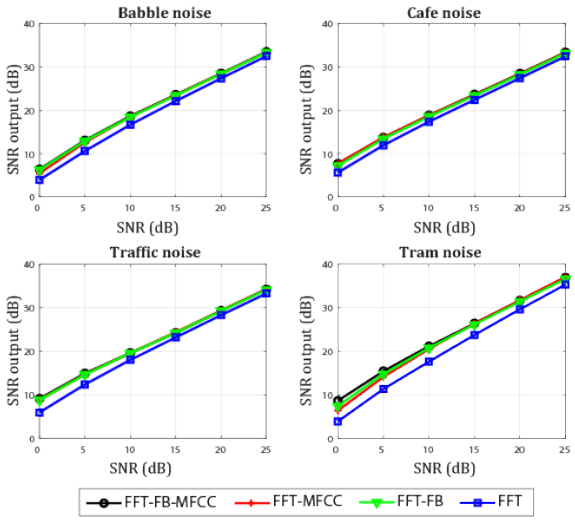


Figure 3: *SNR (dB) vs. SNR output (dB), after enhancement for the DNN-based SE method, with different feature representations.*

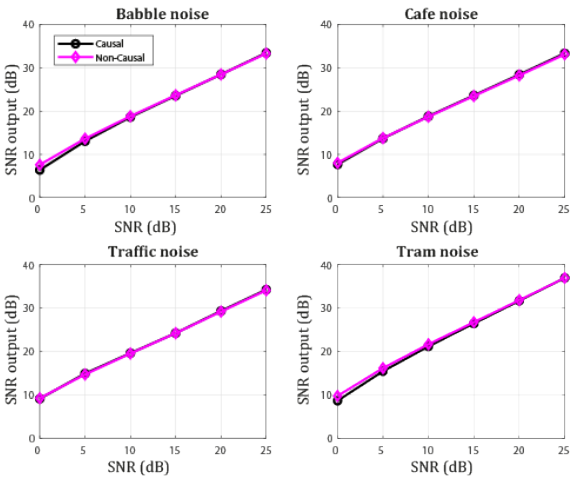


Figure 4: *SNR (dB) vs. SNR output (dB), after enhancement for the DNN-based SE method, using a Causal and Non-causal convolutions.*

Conv1D processing, which maintains the same kernel size ($k = 3$) to define two modalities:

- Causal: uses two previous frames plus the present frame
- Non-causal: uses the previous, the present, and the next frame

Figure 4 presents results of the SNR level using Causal and Non-causal designs. Note that both results are mostly the same for all noise types and SNR_{input} levels evaluated. This result is not surprising if we consider the small size of the kernel. With $k = 3$ there is not a remarkable difference between the segment used for Causal vs. Non-causal convolutions. Anyway this result provides us with the evidence for proposing the use of a causal design on top of a non-causal design, in order to have the option of performing an online enhancement.

5.3. Residual connections

The incorporation of residual connections brought more potential to the CNN approach. They make use of shortcut connections between neural network layers, allowing to handle deeper and more complicated neural network architectures, with fast convergence and a small gradient vanishing effect. Thus, they are more expressive and provide more detailed representations of the underlying structure of the corrupted signal, resulting in more accurate enhanced speech. Residual connections have been previously used for speech enhancement in the form of Wide Residual Network (WRN) architecture [34, 35, 32, 33].

In the following, we compare the CNN architecture with a version of it that includes residual connections (ResNet). Figure 5 shows the SNR level results for both architectures CNN and ResNet. Note that consistently with the conclusion of previous works, results show that ResNet architecture achieves better SE performance, in terms of higher SNR, than the ones obtained using CNN architecture, for all the noise types evaluated.

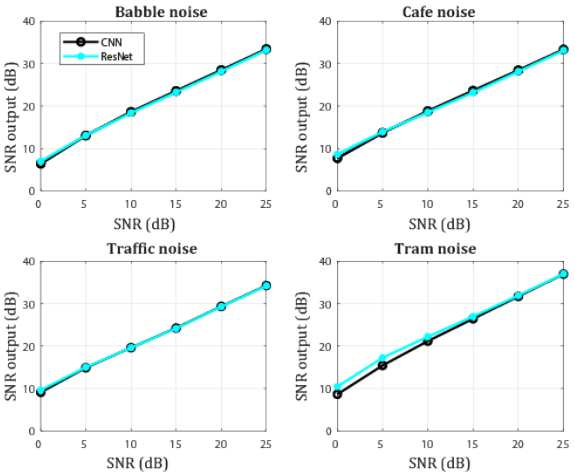


Figure 5: *SNR (dB) vs. SNR output (dB), after enhancement for the DNN-based SE method, using CNN and ResNet architectures.*

5.4. Summary: DNN architecture for Wiener gain estimation

In the following, table 1 presents a summary of obtained results on the assessment of feature representations, residual connections, and causality. This summary provides conclusions that contribute to highlight good practices to perform speech enhancement with DNN. From table 1 we can conclude that the combination that achieves the best performance among the evaluated ones is FFT-FB-MFCC speech features, with residual connections, and non-causal convolutions. Besides the average of the SNR, we also provide Log-Likelihood Ratio as a metric of distortion [36]. LLR represents the degree of discrepancy between smoothed spectra of the target and reference signals, computed over the active speech segments of the Linear Prediction Coefficients.

Issue	**Baseline**	ΔSNR	*LLR*
Feature Representations	FFT	7.17	0.52
	FFT-FB	8.88	0.54
	FFT-MFCC	8.89	0.57
	FFT-FB-MFCC	9.16	0.55
Online processing	Causal	9.16	0.55
	Non-Causal	9.33	0.54
Residual connections	CNN	9.16	0.55
	ResNet	9.38	0.58

Table 1: Average through SNR input level (0-25 dB) and different noise types (babble, tram, traffic, cafe) for $\Delta SNR = SNR_{output} - SNR_{input}$ (dB) and *LLR* for the issues at first column on the DNN-based SE configuration.

So, from the findings on the study performed in this section and the summary of results in table 1 we finally designed the DNN architecture for estimating G_W. It consists of a Wide-ResNet (WRN) architecture with multiple speech representations as input, based on 1-dimensional causal convolutions. Note that despite results indicated that non-causal convolutions achieved better results, these were not so much different. So we decided to have the availability of online processing keeping the causal convolutions.

6. Performance of speech enhancement method

In this section we assessed the performance of the speech enhancement method proposed in section 3 through speech quality measures. This time we estimate the SNR output level using WADA [37], and we measure the quality using PESQ: Perceptual evaluation of speech quality [38] (from 0.5 to 4.5). In both cases, larger values indicate better speech quality.

6.1. Results and discussion

Figure 6 shows the output SNR vs. input SNR for enhanced speech in different *noisy* conditions. The black line is the reference that corresponds to the noisy speech, in this case the SNR estimation algorithm was directly applied to the noisy speech without enhancement. The blue line represents the statistical-based *OMLSA* method, and the red line is related to the *proposal*. For comparison purposes, the green line corresponds to a state-of-the-art DNN-based SE method *SEGAN* [39].

First of all, we observe that the SNR estimation method, WADA, over-estimates slightly at high SNR level. See the corresponding SNR estimation of the noisy speech from 15 dB. However, in practice, at this SNR level this is not a big issue, because either speech with SNR around 20 dB is not usually submitted to speech enhancement. On the other side, all data enhanced with *OMLSA*, *SEGAN*, and *proposal* show a noticeable increment of the SNR with respect to the noisy speech. However, considering some difference in the numerical results, for all noise types the proposal obtains the highest performance.

PESQ results are presented in figure 7. In line with the previous obtained results the upper curve of PESQ values correspond to the proposal, indicating the best performance in terms of audio quality is corresponding to the proposal. This achievement is consistent for the four noise types evaluated.

7. Conclusions

This paper has explored the mutually beneficial relationship between the traditional gain-based approach for speech enhancement and the deep learning solution. The proposal focuses on improving the performance of the Wiener gain estimator by using a DNN architecture designed for SE purposes. We presented a DNN-based SE method in the framework of the OMLSA speech estimation algorithm. Experimental results show that SNR and PESQ values in the state-of-the-art. Beyond these, audio examples of simulated and real noisy speech demonstrate the accuracy of the proposal in practice [22].

During the design of a DNN architecture, we studied various configuration issues frequently related to the DNN-based SE solutions. First, we addressed the use of complementary speech representations for the input of the DNN. Obtained results support the advantage of using them on top of a single spectral parametrization. Then, about the use of residual connections, we conclude that the residual mechanism was beneficial in this

Figure 6: SNR initial vs. output for enhanced and noisy speech.

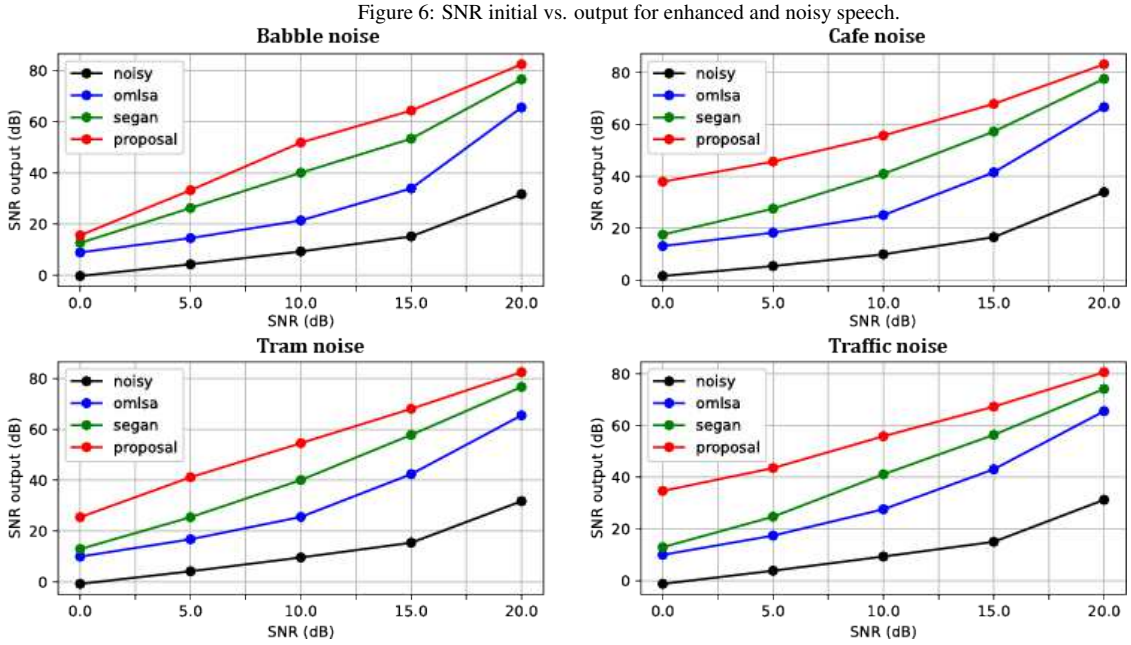
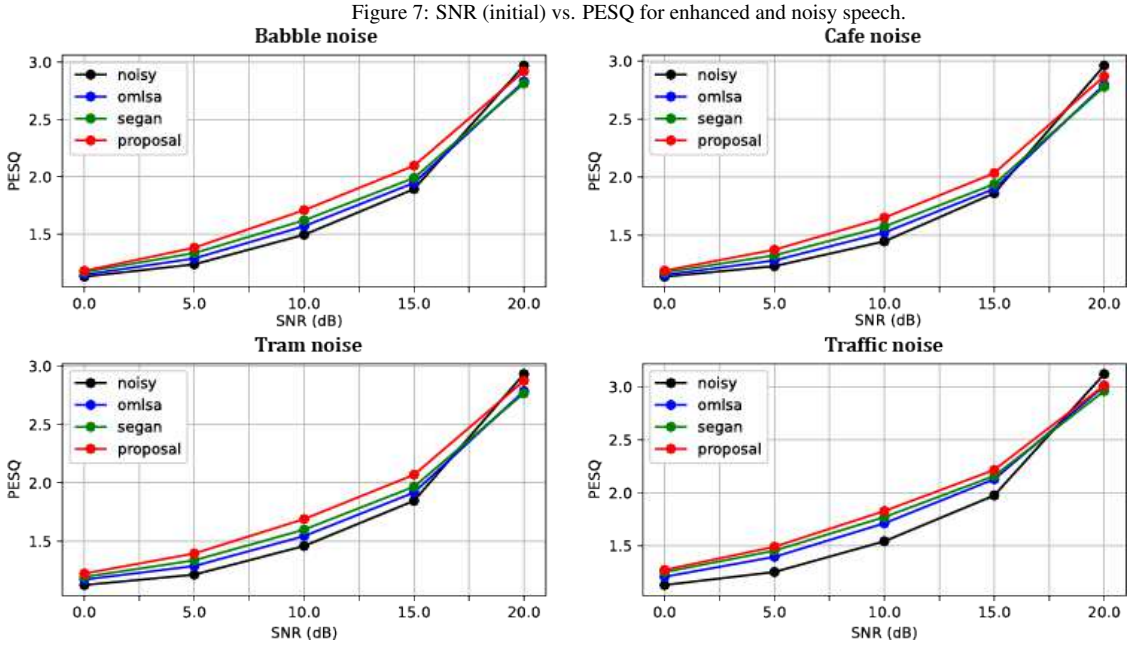


Figure 7: SNR (initial) vs. PESQ for enhanced and noisy speech.



case, supporting the initial motivation of using ResNet for SE. However, it is probably that more complex and deeper architectures can take better advantage of the update provided by the residual connections, giving place to more contrasting results.

Next steps will include testing further complementary parametrization such as the derivatives of MFCC that contributes with speech suprasegmental information. Furthermore we plan to evaluate the performance using causal vs. non-causal architectures including a wider range of kernel size. This last is motivated by the applicability of causal approaches in online applications such as speech enhancement in telephone audio streams.

Abbreviations

Adam: Adaptive Moment Estimator; **ASR**: Automatic Speech Recognition; **CNN**: Convolutional Neural Network; **DNN**: Deep Neural Network; **DSE**: Deep Speech Enhancement; G_{MMSE}: Gain of the Minimum Mean Square Error Estimator; **IBM**: Ideal Binary Mask; **IRM**: Ideal Ratio Mask; **LSA**: Log-Spectral Amplitude; **MMSE**: Minimum Mean Square Error; **OMLSA**: Optimal Modified Log-Spectral Amplitude; **PReLU**: Parametric Rectified Linear Unit; **PSD**: Power Spectral Density; **SNR**: Signal-to-Noise Ratio; **SS**: Spectral Subtraction; **STD**: Standard Deviation; **STFT**: Short-Term Fourier Transform; **STSA**: Short-Time Spectral Amplitude.

Availability of data and materials

Enhanced audio samples are available at: `http://dayanaribas.vivolab.es/DEMOenhancement/index.html`.

About data, Librispeech and Musan datasets supporting the conclusions of this article are available in the openslr repository, `http://openslr.org`. Also Timit dataset is available in LDC repository, `https://catalog.ldc.upenn.edu/LDC93S1W`.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Dayana Ribas, Aantonio Miguel, and Alfonso Ortega designed the proposal. Dayana Ribas performed the set of experiments, and wrote the manuscript. AM designed the DNN architecture. Alfonso Ortega contributes on the manuscript writing and guided the analysis of the results provided in this manuscript. Eduardo Lleida helped to revise the manuscript and approved it for publication. The final manuscript was read and approved by all the authors.

Funding

This work has been supported by System One NOC & Development Solutions S.A. (SONOC) and the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the project TIN2017-85854-C4-1-R, by the Government of Aragon (Reference Group T3617R) and co-financed with Feder 2014-2020 "Building Europe from Aragón".

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research. This material is based upon work supported by Google Cloud.

References

[1] Loizou, P.C.: Speech Enhancement: Theory and Practice. CRC Press, New York (2013)

[2] Hendriks, R.C., Gerkmann, T., Jensen, J.: DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art. Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool, New York (2013)

[3] Lim, J.S., Oppenheim, A.V.: Enhancement and bandwidth compression of noisy speech. Proceedings of the IEEE **67**(12), 1586–1604 (1979)

[4] Boll, S.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. on Acoustic, Speech and Signal Processing **27**(2), 113–120 (1979)

[5] Ephraim, Y., Malah, D.: Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. on Acoustic, Speech and Signal Processing **32**(6), 1109–1121 (1984)

[6] Ephraim, Y., Malah, D.: Speech enhancement using minimum-mean square log spectral amplitude estimator. IEEE Trans. on Acoustic, Speech and Signal Processing **33**(2), 443–445 (1985)

[7] Breithaupt, C., Martin, R.: Analysis of the Decision-Directed SNR Estimator for Speech Enhancement with Respect to Low-SNR and Transient Conditions. IEEE Trans. Speech and Audio Processing (2010)

[8] Xia, B.-Y., Bao, C.-C.: Speech enhancement with weighted denoising auto-encoder. In: Proc. Interspeech (2013)

[9] Xia, B.-Y., Bao, C.-C.: Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. Speech Communication **60**, 13–29 (2014)

[10] Wang, D., Chen, J.: Binary and ratio time-frequency masks for-robust speech recognition. Speech Communication **48**, 1486–1501 (2006)

[11] Wang, D., Chen, J.: Supervised speech separation based on deep learning: An overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing **26**, 1702–1726 (2018)

[12] Narayanan, A., Wang, D.L.: Ideal ratio mask estimation using deep neural networks for robust speech recognition. In: IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), pp. 7092–7096 (2013)

[13] Narayanan, A., Wang, D.L.: Investigation of speech separation as a front-end for noise robust speech recognition. IEEE Transactions on Audio, Speech and Language Processing **22**(4), 826–835 (2014)

[14] Healy, E.W., Yoho, S.E., Wang, Y., Wang, D.: An algorithm to improve speech recognition in noise for hearing-impaired listeners. The Journal of the Acoustical Society of America **134**(4), 3029–3038 (2013)

[15] Healy, E.W., Yoho, S.E., Wang, Y., Apoux, F., Wang, D.: Speech-cue transmission by an algorithm to increase consonant recognition in noise for hearing-impaired listeners. The Journal of the Acoustical Society of America **136**(6), 3325–3336 (2014)

[16] Healy, E.W., Yoho, S.E., Chen, J., Wang, Y., Wang, D.: An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type. The Journal of the Acoustical Society of America **138**(3), 1660–1669 (2015)

[17] Healy, E.W., Delfarah, M., Johnson, E.M., Wang, D.: A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation. The Journal of the Acoustical Society of America **145**, 1378–1388 (2019)
[18] Bolner, F., Goehring, T., Monaghan, J.J., van Dijk, B., Wouters, J., Bleeck, S.: Speech enhancement based on neural networks applied to cochlear implant coding strategies. In: ICASSP, pp. 6520–6524 (2016)
[19] Goehring, T., Bolner, F., Monaghan, J.J., van Dijk, B., Zarowski, A., Bleeck, S.: Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users. The Journal of Hearing research **344**, 183–194 (2017)
[20] Goehring, T., Keshavarzi, M., Carlyon, R.P., Moore, B.C.J.: Using recurrent neural networks to improve the perception of speech in non-stationary noise by people with cochlear implants. The Journal of the Acoustical Society of America **146**(1), 705–708 (2019)
[21] Cohen, I., Berdugo, B.: Speech enhancement for non-stationary noise environments. Signal Processing **81**(11), 2403–2418 (2001)
[22] `http://dayanaribas.vivolab.es/DEMOenhancement/index.html`
[23] McAulay, R.J., Malpass, M.L.: Speech enhancement using a soft-decision noise supression filter. IEEE Transactions on Acoustics, Speech, and Signal Processing **28**(2), 137–145 (1980)
[24] Malah, D., Cox, R.V., Accardi, A.J.: Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (1999)
[25] Hirsch, H., Ehrlicher, C.: Noise estimation techniques for robust speech recognition. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 153–156 (1995)
[26] Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech and Audio Processing **9**, 504–512 (2001)
[27] Snyder, D., Chen, G., Povey, D.: MUSAN: A Music, Speech, and Noise Corpus. arXiv:1510.08484v1 (2015). 1510.08484
[28] Ortega, A., Sukno, F., Lleida, E., Frangi, A., Miguel, A., Buera, L., Zacur, E.: AV@CAR: a Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition. In: Language Resources and Evaluation (LREC), pp. 763–766 (2004)
[29] Kingma, D.P., Ba, J.L.: Adam: Amethod for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR) (2015)
[30] Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam. arXiv preprint arXiv:1711.05101 (2017)
[31] He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
[32] Llombart, J., Ribas, D., Miguel, A., Vicente, L., Ortega, A., Lleida, E.: Speech enhancement with wide residual networks in reverberant environments. In: Interspeech, pp. 1811–1815 (2019)
[33] Llombart, J., Ribas, D., Miguel, A., Vicente, L., Ortega, A., Lleida, E.: Progressive speech enhancement with residual connections. In: Interspeech, pp. 3193–3197 (2019)
[34] Chen, J., Wang, D.: Long short-term memory for speaker generalization in supervised speech separation. The Journal of the Acoustical Society of America **141**(6), 4705–4714 (2017)
[35] Tu, M., Zhang, X.: Speech enhancement based on deep neural networks with skip connections. In: IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), pp. 5565–5569 (2017)
[36] Loizou, P.C.: Speech Quality Asssessment. In: Multimedia Analysis, Processing and Communications, pp. 623–654. Springer, ??? (2011)
[37] Chanwoo Kim, R.M.S.: Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In: Interspeech, pp. 2598–2601 (2008)
[38] 862, R.I.-T.P.: TU-T 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 (2001)
[39] Pascual, S., Bonafonte, A., Serr, J.: Segan: Speech enhancement generative adversarial network. In: INTERSPEECH, pp. 3642–3646 (2017)