Peer-reviewed version available at *Biomimetics* **2019**, *5*, 1; doi:10.3390/biomimetics5010001

Article

Evaluation of Mixed Deep Neural Networks for Reverberant Speech Enhancement

Michelle Gutiérrez-Muñoz [1,†,‡], Astryd González-Salazar [1,‡] and Marvin Coto-Jiménez [1,*]

[1] Affiliation 1; michelle.gutierrezmunoz, astryd.gonzalez,marvin.coto@ucr.ac.cr
* Correspondence: marvin.coto@ucr.ac.cr
† Current address: Escuela de Ingeniería Eléctrica, Universidad de Costa Rica
‡ These authors contributed equally to this work.

Abstract: Speech signals are degraded in real life environments, product of background noise or reverberation. The processing of such signals for voice recognition and voice analysis systems present important challenges. One of the conditions that represent adverse quality difficult to handle in those systems are reverberation, produced by the sound wave reflections that travel from the source to the microphone in multiple directions.To enhance signals in such adverse condition, several Deep Learning-based methods have been proposed and proven to be effective. Recently, recurrent neural networks, especially those with short and long term memory (LSTM), have presented surprising results in tasks related to time-dependent processing of signals, such as the speech. One of the most challenging aspects of LSTM networks is the high computational cost of the training procedure, which have represented a limitation for extended experimentation in several references. In this work, we present a proposal to evaluate the hybrid models of neural networks to learn different reverberation conditions without any previous information. The results show that some combination of LSTM and perceptron layers produce good results in comparison to those of pure LSTM networks. The evaluation has been made based on quality measurements of the signal's spectrum, training time of the networks and statistical validation of results. Results help to affirm the fact that hybrid networks represent an important alternative to this tasks with advantages in efficiency without quality drop.

Keywords: artificial neural network, deep learning, LSTM, speech processing.

1. Introduction

In real environments, audio signals are affected by conditions such as additive noise, reverberation, and other distortions, due to elements that produce sounds simultaneously or are presented as obstacles in the signal path to the microphone. In the case of speech signals, communication devices and applications of speech technologies may be affected in their performance [1–4] in the presence of such conditions.

In the last decades, many algorithms have been developed and to enhance degraded speech, which tries to suppress or reduce the distortions, as well as preserve or improve the quality of the perceived signal [5]. A considerable number of recent algorithms are based on deep neural networks (DNN) [6–9]. The most common implementation is based on approximating a mapping function from the degraded characteristics of speech with noise, towards the corresponding characteristics of clean speech.

The benefits of achieving this type of speech signal enhancement can be applied to signal processing in mobile phone applications, voice over Internet protocol, speech recognition systems and devices for people with a decrease in their hearing ability [10].

In addition to the classical perceptron model, created in the 1950s, new types of neural networks have been developed, for example contemplating recurring connections (RNNs). One of the recent types has been Long Short-Term Memory (LSTM) neural networks. In previous references, to enhance speech, spectrum-derived characteristics, such as Mel-frequency Cepstrum Coefficients (MFCC), have been mapped successfully between clean speech to clean speech [11,12].

The benefits of using LSTM, as well as other types of RNNs, are the best modeling of the dependent nature in speech signals. Among its drawbacks is the high computational cost of its training procedures.

In this work, we extend the previous experiences of experimentation with LSTM by evaluating deep neural networks, with three hidden layers, that combine LSTM layers (bidirectional) and simpler layers, based on perceptrons.

Such type of deep neural network algorithms have been successful in overcoming the performance of classical methods based on signal processing, which have considered various signal-to-noise (SNR) [12–15], or reverberant speech [16–18]. Some recent work has explored the use of Mixed Neural Networks to achieve a better performance in different tasks, such as classifying the temporary stages of sleep, analyzing the real-time behavior of an online buyer or the suppression of noise in a MEMS gyroscope, in which good results were obtained for specific situations and configurations [19], [20], [21].

In our case, the focus is mainly on efficiency in performing the task of interest. To assess the efficiency, we consider different combinations of layers for de-reverberation, intending to accelerate the training process. We intend to measure the ability of LSTM networks to improve voice signals without prior information on the degradation of the signals.

For this purpose, several objective measures are used to verify the results, which comparatively show the capacity of the LSTM with three layers, and the combination with layers of perception, in improving speech conditions of reverberation. The rest of this document is organized as follows: the Section 2 provides the background and context of the problem of improving reverberant speech and the LSTM, the Section 4 describes the experimental setup, the Section 5 presents the results with a discussion, and finally, in the Section 6 conclusions are presented.

2. Problem statement

In real-world environments where speech signals are registered with microphones, the presence of reverberation is common, which is caused by the reflections of the audio signal in its path to the microphone.

This phenomenon is accentuated when the space is wide and the surfaces favor the reflection of the signals. It can be assumed that the reverberated signal x is a degraded version of the clean signal s. The relationship between both waves is described by [22]:

$$x(n) = \mathbf{h}^T(n) * \mathbf{s}(n), \tag{1}$$

where $\mathbf{h} = [h_1, h_2, \ldots, h_L]^T$ is the impulse response of the acoustic channel from the source to the microphone, and $*$ the convolution operation.

The degraded speech signal with reverberation is perceived as distant, as a very short type of echo. Consequently, this effect generally increases as the speaker's distance to the microphone increases.

Since this effect is not desired for proper recognition and analysis of the speech signal, new algorithms have been proposed to minimize it. Mainly, in the last few years, the algorithms based on deep learning have stood out.

By implementing deep neural networks, an approximation to $s(n)$ can be estimated using a function $f(\cdot)$ between the data of the reverberated signal and the clean signal:

$$\hat{s}(t) = f\left(x(t)\right). \tag{2}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2019 doi:10.20944/preprints201910.0376.v1

Peer-reviewed version available at Biomimetics 2019, 5, 1; doi:10.3390/biomimetics5010001

The quality of the approximation performed by $f(\cdot)$ usually depends on the amount of data and the algorithm selected. For the present work, we take as a base case the estimation of $f(\cdot)$ made by BLSTM networks with three hidden layers. In this model, we propose a comparison and statistical validation of results with mixed networks, which include combinations of BLSTM layers and perceptions.

3. Autoencoders of BLSTM networks

Since the appearance of the RNNs, there are new alternatives to model the character dependent on the sequential information in applications where this nature of the parameters is relevant. These types of neural networks are capable of storing information through feedback connections between neurons in their hidden layers or another network that is in the same layer [23,24].

With the purpose of expanding the capabilities of the RNNs by storing information in the short and long term, the LSTM networks shown in [25] introduce a set of gates into the memory cells capable of controlling the access, storage and propagation of values across the network. The results obtained when using LSTM networks in areas that depend on previous states of information, such as the case of voice recognition, musical composition and handwriting synthesis, were encouraging [25–27].

In addition to the recurring connections between the internal units, each unit in the network has additional gates for storing values: an input gate, one for memory clearing, one for output and one for activating memory. In this way, it is possible to store values for many steps, or have them available at any time [25].

The gates are implemented using the following equations:

$$i_t = sigma\left(\mathbf{W}_{xi}x_t + \mathbf{W}_{hi}h_{t-1} + \mathbf{W}_{ci}c_{t-1} + b_i\right) \tag{3}$$

$$f_t = \sigma\left(\mathbf{W}_{xf}x_t + \mathbf{W}_{hf}h_{t-1} + \mathbf{W}_{cf}c_{t-1} + b_f\right) \tag{4}$$

$$c_t = f_t c_{t-1} + i_t \tanh\left(\mathbf{W}_{xc}x_t + \mathbf{W}_{hc}h_{t-1} + b_c\right) \tag{5}$$

$$o_t = \sigma\left(\mathbf{W}_{xo}x_t + \mathbf{W}_{ho}h_{t-1} + \mathbf{W}_{co}c_t + b_o\right) \tag{6}$$

$$h_t = o_t \tanh\left(c_t\right) \tag{7}$$

where σ is the sigmoid activation function, i is the input gate, f the memory erase gate and or the exit gate. c is the activation of memory. $\mathbf{W}_{mn}$ is the matrix that contains the values of the connections between each unit and the gates. h is the output of the LSTM memory unit.

Additional details about the training process and the implications of this implementation can be found at [28].

An additional extension of LSTM networks that has had a greater advantage in tasks related to temporal parameter dependence is the bidirectional LSTM network (BLSTM). In this, the configuration of the network allows the update of parameters in both directions of the process, as if it were not only to convert the input parameters to the reference of the output, but in the opposite direction. In this work, these units will be used to make comparisons.

Training neural networks for the improvement of speech signals and noise reduction became a solid idea from its first application in the correction of binary input patterns. Later, this idea was used in modeling acoustic coefficients, which were mapped using a single layer. The above due to the limitation caused by the capabilities of the computers and the algorithms developed for this purpose at the time [14].

An autoencoder for noise reduction is a neural network architecture that has been successful in various tasks related to speech [29]. This architecture consists of an encoder that transforms an input vector s into a representation in the hidden layers h through a f mapping. It also has a decoder that takes the hidden representation and transforms it back into a vector in the input space.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2019 doi:10.20944/preprints201910.0376.v1

Peer-reviewed version available at Biomimetics 2019, 5, 1; doi:10.3390/biomimetics5010001

During training, the features of the distorted signal (noise or reverberation) are used as inputs of the noise elimination autoencoders, while the features of the clean speech are presented as outputs. In addition, to learn the complex relationships between these sets of features, the training algorithm adjusts the parameters of the network. Currently, computers and algorithms have the ability to process large data sets, as well as networks with several hidden layers.

4. Experimental setup

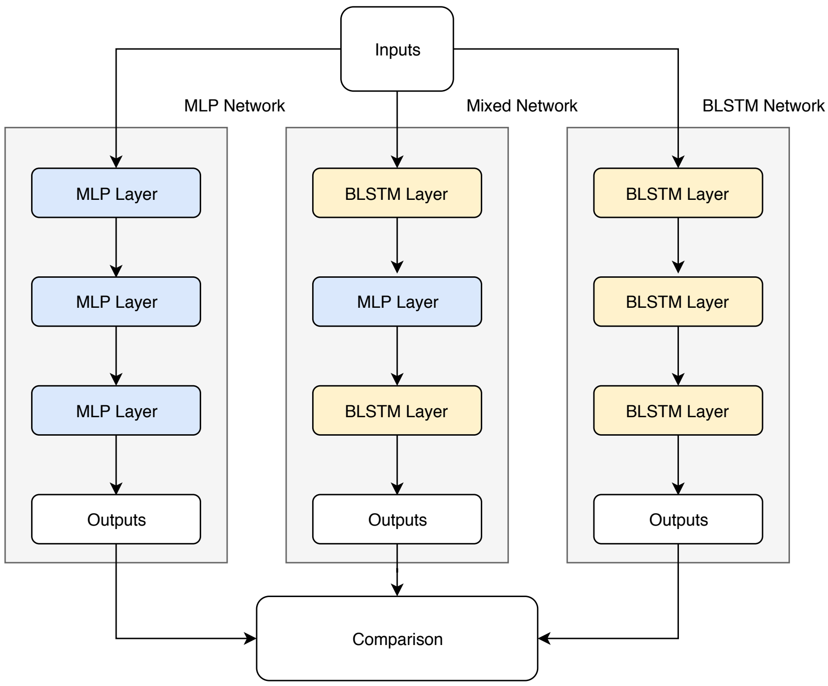

Figure 1. Sample of three networks compared in this work: The purely multi-layer perceptron(MPL), a mixed network, and the purely BLSTM network.

To test our proposed mixed neural networks LSTM / Perceptron to enhance reverberated speech, the experiment can be summarized in the following steps:

1. Selection of conditions: Given the large number of impulse responses contemplated in the databases, we randomly choose five reverberated speech conditions. Each of the conditions has the corresponding clean version in the database.
2. Extraction of features and input-output correspondence: A set of parameters was extracted from the reverberated and clean audio files. Those of the reverberated files were used as inputs to the networks, while the corresponding clean functions were the outputs.
3. Training: During training, the weights of the networks were adjusted as the parameters with reverberation and clean were presented to the network. As usual in recurrent neural networks, the updating of the values of the internal weights is carried out using the back-propagation algorithm through time. A total of 210 expressions were used for each condition (approximately 70 % of the total database) to train each case. The details and equations of the algorithm followed can be found in [30].
4. Validation: after each training step, the sum of the squared errors within the validation set of approximately 20 % of the statements was calculated, and the weights of the network were updated in each improvement.
5. Test: A subset of 50 phrases, selected at random, (about 10 % of the total number of phrases in the database) was chosen for the test set, for each condition. These phrases were not part of the training process, to provide independence between training and testing.

In the following subsections, more details of the experimental procedure are provided.

4.1. Database

In our work, we use the Reverberant Voice Database created at the University of Edinburgh [31], which was designed to train and evaluate the methods of speech de-reverberation. The reverberated speech of the database was produced by convolving the recordings of 56 native English speakers with several impulse responses in various university halls. For this work, we randomly choose the following conditions: ACE Building Lobby 1, Artificial Room 1, Mardy Room 2, ACE Lecture Room 1 and ACE Meeting Room 2.

4.2. Feature extraction

The audio files of the reverberated and clean voice were down-sampled at a rate of 16 kHz, 16 bits, to extract the parameters using the Ahocoder [32] system. A window size of 160 samples and a window shift of 80 samples were used to extract 39 MFCC, f_0 and the energy of each sentence.

For this work, neural networks were applied only to improve the 39 MFCC coefficients, while the rest of the parameters remained invariant.

4.3. Evaluation

For the evaluation of the results, the following objective measures were applied:

- Perceptual evaluation of speech quality (PESQ): This measure uses a model to predict the subjective quality of speech, as defined in ITU-T P.862.ITU recommendation. The results are in the range $[0.5, 4.5]$, where 4.5 corresponds to the signal enhanced perfectly. PESQ is calculated as [33]:

$$\text{PESQ} = a_0 + a_1 D_{ind} + a_2 A_{ind} \tag{8}$$

  where D_{ind} is the average disturbance and A_{ind} the asymmetric perturbation. The a_k were chosen to optimize PESQ in the measurement of general speech quality.
- Sum of squared errors (sse): This is the most common metric for the validation set error during the training process of a neural network. It is defined as:

$$\begin{aligned} \text{sse}(\theta) &= \sum_{n=1}^{T} (mathbf{c}_x - \hat{\mathbf{c}_\mathbf{x}})^2 \quad (9) \\ &= \sum_{n=1}^{T} (\mathbf{c}_\mathbf{x} - f(\mathbf{c}_\mathbf{x}))^2 , \quad (10) \end{aligned}$$

  where c_x is the known value of the outputs and $\hat{c_x}$ the approximation made by the network.
- Time per epoch: Refers to the time it takes for an iteration of the training process.

Additionally, Friedman's statistical test has been used to determine the statistical significance of the results in the test sets.

4.4. Experiments

Figure 1 shows the procedure followed for the comparison between the different architectures tested in this work. To analyze all the architectures that can be formed with a mixture of BLSTM layers and MLP layers, a total of eight different neural networks were tested for each reverberation condition:

- BLSTM-BLSTM-BLSTM
- BLSTM - BLSTM - MLP
- BLSTM-MLP-BLSTM

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2019 doi:10.20944/preprints201910.0376.v1

Peer-reviewed version available at *Biomimetics* **2019**, *5*, 1; doi:10.3390/biomimetics5010001

- BLSTM - MLP - MLP
- MLP - BLSTM - BLSTM
- MLP-BLSTM-MLP
- MLP - MLP - BLSTM
- MLP - MLP - MLP

The metrics were applied in each of these possibilities, which constitute all the possibilities that can be combined between the BLSTM and MLP layers in three layers.

5. Results and Discussion

Table 1 shows the training results for all networks and all possible combinations of three hidden layers. The training of each set was repeated or three times, and the average values are reported. By following the reports made in works before this article, the network with only BLSTM layers provides the best results in most cases of reverberation conditions.

Table 1. Efficiency of the different combinations of hidden layers, by the condition of reverberation. * is the best value of sse in each condition

Condition	Network (Hidden layers)	sse	Time per epoch (s)
MARDY	BLSTM-BLSTM-BLSTM	201.34*	50.6
	BLSTM - BLSTM - MLP	204.39	33.3
	BLSTM-MLP-BLSTM	210.81	33.5
	BLSTM - MLP - MLP	218.91	15.9
	MLP - BLSTM - BLSTM	204.82	36.1
	MLP-BLSTM-MLP	256.32	18.6
	MLP - MLP - BLSTM	216.46	18.8
	MLP - MLP - MLP	400.34	1.2
Lecture Room	BLSTM-BLSTM-BLSTM	213.12	74.9
	BLSTM - BLSTM -MLP	214.35	48.8
	BLSTM-MLP-BLSTM	221.88	49.3
	BLSTM - MLP - MLP	229.22	23.2
	MLP - BLSTM - BLSTM	212.34*	52.8
	MLP-BLSTM-MLP	226.39	27.7
	MLP - MLP -BLSTM	230.85	27.6
	MLP-MLP-MLP	360.41	1.8
Artificial Room	BLSTM-BLSTM-BLSTM	88.47*	55.5
	BLSTM - BLSTM -MLP	90.37	36.5
	BLSTM-MLP-BLSTM	93.61	36.6
	BLSTM - MLP - MLP	104.23	17.4
	MLP - BLSTM - BLSTM	92.18	39.5
	MLP-BLSTM-MLP	108.56	20.6
	MLP - MLP -BLSTM	111.13	20.5
	MLP-MLP-MLP	170.61	1.3
ACE Building	BLSTM-BLSTM-BLSTM	207.32*	73.8
	BLSTM - BLSTM -MLP	210.17	45.8
	BLSTM-MLP-BLSTM	214.29	46.1
	BLSTM - MLP - MLP	212.54	21.6
	MLP - BLSTM - BLSTM	208.04	49.2
	MLP-BLSTM-MLP	221.28	25.6
	MLP - MLP -BLSTM	220.13	25.8
	MLP-MLP-MLP	333.60	1.7
Meeting Room	BLSTM-BLSTM-BLSTM	197.37	69.9
	BLSTM - BLSTM -MLP	199.03	45.7
	BLSTM-MLP-BLSTM	204.68	45.8
	BLSTM - MLP - MLP	217.52	21.6
	MLP - BLSTM - BLSTM	196.90*	49.6
	MLP-BLSTM-MLP	206.03	25.7
	MLP - MLP -BLSTM	214.28	25.9
	MLP-MLP-MLP	363.19	1.7

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 October 2019 doi:10.20944/preprints201910.0376.v1

Peer-reviewed version available at *Biomimetics* **2019**, *5*, 1; doi:10.3390/biomimetics5010001

Table 2. Objective evaluations for the different combinations of hidden layers, by the condition of reverberation. * is the best value. The p-value was obtained with the Friedman test, with a significance of 0.05.

Condition	Network (Hidden layers)	PESQ	Significative difference	p-value
MARDY	BLSTM-BLSTM-BLSTM	2.30	-	-
	BLSTM - BLSTM - MLP	2.31*	no	0.715
	BLSTM-MLP-BLSTM	2.27	yes	0.003
	BLSTM - MLP - MLP	2.19	yes	6.648e-08
	MLP - BLSTM - BLSTM	2.28	no	0.147
	MLP-BLSTM-MLP	2.08	yes	1.965e-14
	MLP - MLP - BLSTM	2.24	yes	0.000
	MLP - MLP - MLP	1.94	yes	0.000
Lecture Room	BLSTM-BLSTM-BLSTM	2.28*	-	-
	BLSTM - BLSTM - MLP	2.21	no	0.095
	BLSTM-MLP-BLSTM	2.22	yes	0.0034
	BLSTM - MLP - MLP	2.20	yes	1.729e-07
	MLP - BLSTM - BLSTM	2.27	no	0.199
	MLP-BLSTM-MLP	2.21	yes	9.635e-05
	MLP - MLP - BLSTM	2.20	yes	9.617
	MLP - MLP - MLP	2.00	yes	0.000
Artificial Room	BLSTM-BLSTM-BLSTM	3.18*	-	-
	BLSTM - BLSTM - MLP	3.17	no	1.000
	BLSTM-MLP-BLSTM	3.14	yes	0.002
	BLSTM - MLP - MLP	3.12	yes	6.650e-08
	MLP - BLSTM - BLSTM	3.17	no	1.000
	MLP-BLSTM-MLP	3.06	yes	1.965e-14
	MLP - MLP - BLSTM	3.08	yes	2.695e-06
	MLP - MLP - MLP	2.90	yes	0.000
ACE Building	BLSTM-BLSTM-BLSTM	2.37*	-	-
	BLSTM - BLSTM - MLP	2.35	no	0.068
	BLSTM-MLP-BLSTM	2.35	no	0.147
	BLSTM - MLP - MLP	2.32	yes	4.22e-05
	MLP - BLSTM - BLSTM	2.36	no	0.474
	MLP-BLSTM-MLP	2.33	yes	0.026
	MLP - MLP - BLSTM	2.33	yes	0.008
	MLP - MLP - MLP	2.08	yes	0.000
Meeting Room	BLSTM-BLSTM-BLSTM	2.28	-	-
	BLSTM - BLSTM - MLP	2.29*	no	0.147
	BLSTM-MLP-BLSTM	2.24	no	0.060
	BLSTM - MLP - MLP	2.23	yes	0.002
	MLP - BLSTM - BLSTM	2.28	no	0.474
	MLP-BLSTM-MLP	2.25	no	0.715
	MLP - MLP - BLSTM	2.20	yes	0.001
	MLP - MLP - MLP	2.0	yes	1.960e-14

For the five cases of reverberation considered in this paper, the network that stands out as a competitive alternative to the three-layer BLSTM network is the MLP-BLSTM-BLSTM configuration. In addition to presenting in two cases a better result between all the architectures (under the conditions "Lecture Room" and "Meeting Room”), the training time is almost 30% less per epoch in Comparison to the BLSTM network. This is one of the main indicators sought in this work.

In the same Table 1, it is seen how the training times are similar between those configurations consisting of two BLSTM layers and one MLP, ayes as between those of only one BLSTM layer and two MLP. The MLP-MLP-MLP type networks, despite having very low training times per season, as expected, do not present competitive results in comparison to others.

In addition to the verification regarding the training efficiency of the networks, Table 2 shows the results in terms of the PESQ quality metric. This is of the utmost importance since the analysis of the problem of de-reverberation of speech signals is being raised. So improvements in efficiency and sse values must also be checked in terms of the quality of the signal achieved.

In these last table, the differences obtained for the BLSTM-BLSTM-BLSTM base system are presented, in terms of statistical significance according to the Friedman test.

In each of the five reverberation conditions, the results of these tests can be summarized:

- MARDY, Lecture Room and Artificial Room: Only two of the mixed configurations present results that do not differ statistically significantly with the base system. These mixed networks are BLSTM-BLSTM-MLP and MLP-BLSTM-BLSTM.
- Ace Building: In this case, three combinations of hidden layers present results that do not differ significantly from the base case.
- Meeting Room: This is a particular case, because in the combination BLSTM-BLSTM-MLP is the one that presents the best result, although the improvement is not significant compared to the base system. On the other hand, both MLP-BLSTM-BLSTM and BLSTM-MLP-BLSTM and MLP-BLSTM-MLP present results that do not differ significantly.

In the Figure 2 it can be seen the spectrograms corresponding to clean speech, to speech with reverberation and to two of the proposed configurations: That based solely on BLSTM layers, and the mixed network that obtained better results (MLP-BLSTM-BLSTM). It is possible to appreciate the improvements introduced by the neural networks and the proximity that is perceived visually in this representation between the spectrogram of the mixed network in comparison to the base system.

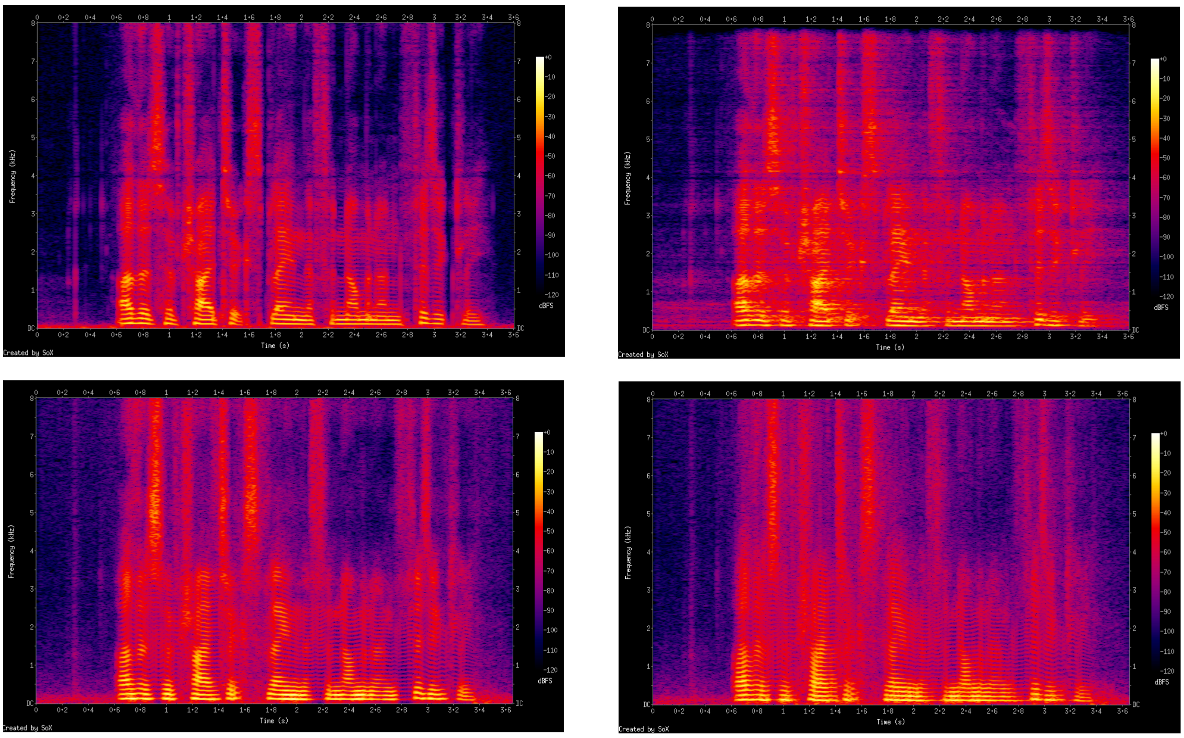

Figure 2. Spectrograms of a phrase in the database. Upper left: speak clean. Top right: Speak with reverberation (ACE Building Lobby). Bottom left: Enhancement result with the BLSTM network. Bottom right: Enhancement result with the mixed MLP-BLSTM-BLSTM network.

Considering the previous efficiency results and how these are reflected in the PESQ metric, it is emphasized that there are combinations of mixed networks, especially MLP-BLSTM-BLSTM, which reduce the times of training considerably, without significantly sacrificing the quality of results in the reverberation of the signals.

Peer-reviewed version available at *Biomimetics* **2019**, *5*, 1; doi:10.3390/biomimetics5010001

6. Conclusions

In this work, the use of mixed neural networks, consisting of combinations of layers formed by perceptron units, with BLSTM layers, was proposed as an alternative for the reduction of training time of purely BLSTM networks. Training time has represented a limitation for extensive experimentation with this type of artificial neural networks in different applications, including some related to the improvement of speech signals.

One of the eight possible combinations of mixed networks presented competitive results in terms of the metrics of the training system and results that do not differ significantly from the purely BLSTM case in terms of PESQ of the signals. The significance was determined with a statistical test. The reduction in training time is of the order of 30 %, in processes that can normally take hours or days, depending on the amount of data.

The results presented open the possibility of simplifying some neural network configurations to be able to perform extensive experimentation in different applications where it is required to map parameters of such nature, as in the case of autoencoders.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing–original draft preparation, X.X.; writing–review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: This research received no external funding

Acknowledgments: This work was made with the support of the University of Costa Rica, project 322-B9-105.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
TLA	Three letter acronym
LD	linear dichroism

References

1. Weninger, F.; Watanabe, S.; Tachioka, Y.; Schuller, B. Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition. Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014, pp. 4623–4627.
2. Weninger, F.; Geiger, J.; Wöllmer, M.; Schuller, B.; Rigoll, G. Feature enhancement by deep LSTM networks for ASR in reverberant multisource environments. *Computer Speech & Language* **2014**, *28*, 888–902.
3. Narayanan, A.; Wang, D. Ideal ratio mask estimation using deep neural networks for robust speech recognition. Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013, pp. 7092–7096.
4. Bagchi, D.; Mandel, M.I.; Wang, Z.; He, Y.; Plummer, A.; Fosler-Lussier, E. Combining spectral feature mapping and multi-channel model-based source separation for noise-robust automatic speech recognition. Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on. IEEE, 2015, pp. 496–503.
5. Hansen, J.H.; Pellom, B.L. An effective quality evaluation protocol for speech enhancement algorithms. Fifth International Conference on Spoken Language Processing, 1998.
6. Du, J.; Wang, Q.; Gao, T.; Xu, Y.; Dai, L.R.; Lee, C.H. Robust speech recognition with speech enhanced deep neural networks. Fifteenth Annual Conference of the International Speech Communication Association, 2014.

7. Han, K.; He, Y.; Bagchi, D.; Fosler-Lussier, E.; Wang, D. Deep neural network based spectral feature mapping for robust speech recognition. Sixteenth Annual Conference of the International Speech Communication Association, 2015.
8. Maas, A.L.; Le, Q.V.; O'Neil, T.M.; Vinyals, O.; Nguyen, P.; Ng, A.Y. Recurrent neural networks for noise reduction in robust ASR. Thirteenth Annual Conference of the International Speech Communication Association, 2012.
9. Deng, L.; Li, J.; Huang, J.T.; Yao, K.; Yu, D.; Seide, F.; Seltzer, M.L.; Zweig, G.; He, X.; Williams, J.D.; others. Recent advances in deep learning for speech research at Microsoft. ICASSP, 2013, Vol. 26, p. 64.
10. 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* **2013**, *134*, 3029–3038.
11. Coto-Jiménez, M.; Goddard-Close, J. LSTM Deep Neural Networks Postfiltering for Enhancing Synthetic Voices. *International Journal of Pattern Recognition and Artificial Intelligence* **2018**, *32*, 1860008.
12. Coto-Jiménez, M. Robustness of LSTM Neural Networks for the Enhancement of Spectral Parameters in Noisy Speech Signals. Mexican International Conference on Artificial Intelligence. Springer, 2018, pp. 227–238.
13. Kumar, A.; Florencio, D. Speech enhancement in multiple-noise conditions using deep neural networks. *arXiv preprint arXiv:1605.02427* **2016**.
14. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.r.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N.; others. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine* **2012**, *29*, 82–97.
15. Vincent, E.; Watanabe, S.; Nugraha, A.A.; Barker, J.; Marxer, R. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language* **2017**, *46*, 535–557.
16. Feng, X.; Zhang, Y.; Glass, J. Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014, pp. 1759–1763.
17. Ishii, T.; Komiyama, H.; Shinozaki, T.; Horiuchi, Y.; Kuroiwa, S. Reverberant speech recognition based on denoising autoencoder. Interspeech, 2013, pp. 3512–3516.
18. Zhao, Y.; Wang, Z.Q.; Wang, D. Two-Stage Deep Learning for Noisy-Reverberant Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2019**, *27*, 53–62.
19. Dong, H.; Supratak, A.; Pan, W.; Wu, C.; M, P.; Matthews.; Guo, Y. Mixed Neural Network Approach for Temporal Sleep Stage Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **2017**, pp. 4–5.
20. Sakar, C.O.; Polat, S.O.; Katircioglu, M.; Kastro, Y. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications* **2018**, pp. 1–16.
21. Jiang, C.; Chen, Y.; Chen, S.; Bo, Y.; Li, W.; Tian, W.; Guo, J. A Mixed Deep Recurrent Neural Network for MEMS Gyroscope Noise Suppressing. *Electronics* **2019**.
22. Naylor, P.A.; Gaubitch, N.D. *Speech dereverberation*; Springer Science & Business Media, 2010.
23. Fan, Y.; Qian, Y.; Xie, F.L.; Soong, F.K. TTS synthesis with bidirectional LSTM based recurrent neural networks. Fifteenth Annual Conference of the International Speech Communication Association, 2014.
24. Zen, H.; Sak, H. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015, pp. 4470–4474.
25. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural computation* **1997**, *9*, 1735–1780.
26. Graves, A.; Jaitly, N.; Mohamed, A.r. Hybrid speech recognition with deep bidirectional LSTM. Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013, pp. 273–278.
27. Graves, A.; Fernández, S.; Schmidhuber, J. Bidirectional LSTM networks for improved phoneme classification and recognition. International Conference on Artificial Neural Networks. Springer, 2005, pp. 799–804.
28. Gers, F.A.; Schraudolph, N.N.; Schmidhuber, J. Learning precise timing with LSTM recurrent networks. *Journal of machine learning research* **2002**, *3*, 115–143.

doi:10.20944/preprints201910.0376.v1

Peer-reviewed version available at *Biomimetics* **2019**, *5*, 1; doi:10.3390/biomimetics5010001

29. Coto-Jimenez, M.; Goddard-Close, J.; Di Persia, L.; Rufiner, H.L. Hybrid Speech Enhancement with Wiener filters and Deep LSTM Denoising Autoencoders. 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI). IEEE, 2018, pp. 1–8.
30. Greff, K.; Srivastava, R.K.; Koutník, J.; Steunebrink, B.R.; Schmidhuber, J. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems* **2017**, *28*, 2222–2232.
31. Valentini-Botinhao, C. Reverberant speech database for training speech dereverberation algorithms and TTS models, 2016.
32. Erro, D.; Sainz, I.; Navas, E.; Hernáez, I. Improved HNM-based vocoder for statistical synthesizers. Twelfth Annual Conference of the International Speech Communication Association, 2011.
33. Rix, A.W.; Hollier, M.P.; Hekstra, A.P.; Beerends, J.G. Perceptual Evaluation of Speech Quality (PESQ) The New ITU Standard for End-to-End Speech Quality Assessment Part I–Time-Delay Compensation. *Journal of the Audio Engineering Society* **2002**, *50*, 755–764.