# Assessing the effectiveness of transfer learning strategies in BLSTM networks for speech denoising

## Evaluación de la eficacia de las estrategias de aprendizaje por transferencia en las redes BLSTM para la reducción del ruido

Marvin Coto-Jiménez[1], Astryd González-Salazar[2], Michelle Gutiérrez-Muñoz[3]

Coto-Jimenez, M., González-Salazar, A., Gutiérrez-Muñoz, M. Assessing the effectiveness of transfer learning strategies in BLSTM networks for speech denoising. *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence Pág. 42-49.

https://doi.org/10.18845/tm.v35i8.6448

[1] Electrical Engineering Department. University of Costa Rica. Costa Rica.
E-mail: marvin.coto@ucr.ac.cr
https://orcid.org/0000-0002-6833-9938

[2] Electrical Engineering Department. University of Costa Rica. Costa Rica.
E-mail: astryd.gonzalez@ucr.ac.cr
https://orcid.org/0000-0002-3444-0464

[3] Electrical Engineering Department. University of Costa Rica. Costa Rica.
E-mail: michelle.gutierrezmunoz@ucr.ac.cr
https://orcid.org/0000-0003-3313-8324

## Keywords

BLSTM; deep learning; speech processing; transfer learning.

## Abstract

Denoising speech signals represent a challenging task due to the increasing number of applications and technologies currently implemented in communication and portable devices. In those applications, challenging environmental conditions such as background noise, reverberation, and other sound artifacts can affect the quality of the signals. As a result, it also impacts the systems for speech recognition, speaker identification, and sound source localization, among many others. For denoising the speech signals degraded with the many kinds and possibly different levels of noise, several algorithms have been proposed during the past decades, with recent proposals based on deep learning presented as state-of-the-art, in particular those based on Long Short-Term Memory Networks (LSTM and Bidirectional-LSMT). In this work, a comparative study on different transfer learning strategies for reducing training time and increase the effectiveness of this kind of network is presented. The reduction in training time is one of the most critical challenges due to the high computational cost of training LSTM and BLSTM. Those strategies arose from the different options to initialize the networks, using clean or noisy information of several types. Results show the convenience of transferring information from a single case of denoising network to the rest, with a significant reduction in training time and denoising capabilities of the BLSTM networks.

## Palabras clave

BLSTM; aprendizaje profundo; procesamiento del habla; aprendizaje por transferencia.

## Resumen

La eliminación de ruido de las señales de voz representa una tarea desafiante debido al creciente número de aplicaciones y tecnologías implementadas actualmente en los dispositivos portátiles y de comunicación. En esas aplicaciones, las condiciones ambientales desafiantes como el ruido de fondo, la reverberación y otros artefactos de sonido pueden afectar la calidad de las señales. Como resultado, también afecta a los sistemas de reconocimiento de voz, identificación de hablantes y localización de fuentes de sonido, entre muchos otros. Para eliminar el ruido de las señales de voz degradadas con los muchos tipos y posiblemente diferentes niveles de ruido, se han propuesto varios algoritmos durante las últimas décadas, con propuestas recientes basadas en el aprendizaje profundo presentadas como vanguardistas, en particular las basadas en redes de memoria a corto plazo (LSTM y LSMT bidireccional). En este trabajo se presenta un estudio comparativo de diferentes estrategias de transferencia de aprendizaje para reducir el tiempo de formación y aumentar la efectividad de este tipo de redes. La reducción del tiempo de entrenamiento es uno de los desafíos más críticos debido al alto costo computacional de entrenar LSTM y BLSTM. Esas estrategias surgieron de las diferentes opciones para inicializar las redes, utilizando información limpia o ruidosa de varios tipos. Los resultados muestran la conveniencia de transferir información de un solo caso de eliminación de ruido de la red al resto, con una reducción significativa en el tiempo de entrenamiento y las capacidades de eliminación de ruido de las redes BLSTM.

## Introduction

In the speech signal processing, clean audio is ideally expected, without additive or convolutional noise. Though, in uncontrolled conditions, audio signals are degraded by multiple unknown agents or conditions that affect the quality of speech. As a result, these conditions do not allow the optimal performance of speech technologies [1–3].

To solve this problem and enhance the signal quality recorded in real-life environments, several algorithms have been developed throughout the years. In this work, a method based on deep neural networks (DNN) was proposed to map the noisy speech to clean speech [4, 5]. As mentioned, with degraded signals, speech technologies do not work properly, for this reason, the DNN approach can be implemented for better results in several applications, such as in mobile phone applications, speech recognition systems, and assistive technology [6, 7].

One example of this successful method of noise reduction is the Long Short- term Memory (LSTM) neural networks and its bidirectional extension (BLSTM), which are models of recurrent neural networks (RNNs) [8]. Particularly in speech recognition, LSTM has shown better results than DNN or convolutional networks [9, 10]. On the order hand, their training procedures represent a high computational cost. For this reason, a study presented in [8] explained the advantages of using mixed neural networks for reducing computational cost in the task of reverberant speech enhancement. In this work, a comparative study is presented on different transfer learning strategies to improve the capacity of BLSTM neural networks for noise reduction and reducing training time in a set of different noise types and levels.

Transfer learning is a concept used in artificial neural networks (ANN), to improve the results of a model in one domain by transferring information from a model in a related domain. This can be described using a given source domain DS with a corresponding task TS, and a target domain DT with a corresponding task TT . The process improves the target predictive function $f_T$ (·) by using the related information from DS and TS [11].

One of its advantages is that it can be applied with a reduced amount of training data in DT [12]. For this work, the homogeneous transfer process was applied, because there is available data that is drawn from a domain (denoising speech degraded with artificial noise) related to, but not an exact match for a target domain of interest (denoising speech degraded with natural noise).

## Problem Statement

The modeling of speech signals degraded with background noise can be presented in its simplest way as the combination of a pure speech signal, x, with an additive noise d. Thus, the noisy signal y(t) can be expressed in the time domain by the sum:

$$y(t)=x(t)+d(t) \tag{1}$$

In discrete frequency-domain, the relation is simply

$$Y_k(n)=X_k(n)+D_k(n) \tag{2}$$

For the statistical implication of several signal processing-based methods, the clean signal x(t) may be modeled as independent and uncorrelated to the noise d(t). This way, the algorithms can attempt to extract the information of $X_k(n)$ from $Y_k(n)$ and $D_k(n)$, and then reconstruct the approximation for x(t).

From the perspective of deep learning approaches that apply artificial neural networks, x(t) (or $X_k(n)$) can be estimated in the form of an approximated function f(.) that is calculated directly from samples of noisy and clean data of the form:

$$\hat{x}(t) = f\,(y(t)) \qquad (3)$$

How well the task is performed by the artificial neural network, such as BLSTM, depends on the amount of available data and the type of network selected [13]. In our work f(.) is obtained not only from the pairs of data y(t) and x(t) but from another function estimated to perform a similar task, with other kinds of noises. This way, transfer learning is performed between two artificial neural networks.

To verify the benefits of this approach, several objective measures were applied, to compare the training of the BLSTM in the task of denoising several types of noise at several SNR levels.

## Experimental setup

The experimental setup to test the Transfer Learning strategies in BLSTM Networks can be condensed in three steps:

1. Speech dataset generation: Two kinds of noise, one artificially generated (White Noise), and one naturalistic (the crowd noise Babble) were added to each clean utterance recorded in the dataset for different signal-to-noise ratio (SNR -10 dB, SNR -5 dB, SNR 0 dB, SNR 5 dB and SNR 10 dB), with the aim of covering light to heavy noise distortion in White and Babble noise.
2. Feature extraction: Owing to previous experiences in denoising autoencoders,MFCC information was extracted from each frame of the clean utterances and those degraded with noise, using the Ahocoder system [14]. Thus, pairs of noisy and clean parameters were presented to the networks as inputs and outputs during training.
3. Training and testing: The BLSTM networks are trained using backpropagation through time algorithm, to adjust the internal set of values in the connections according to the pairs of values presented to each BLSTM network. The total database of about 900 utterances was splited 80% for training, 15% for validation and 5% for the test (about 50 files). Details and equations of the training algorithm followed can be found in [15].

### Dataset

SLT (female) was selected from the CMU Arctic database [16], widely applied in speech research. This dataset were designed and produced at the Carnegie Mellon University. The SLT set of utterances consists of 1132 sentences, and split the whole set randomly to establish the training, validation, and test set, as described in the previous section.

### Transfer Learning Strategies

One base system (training without transferring information) and three Transfer Learning strategies were compared:

- *Case 1 (Base system):* This case corresponds to the traditional random initialization of the internal weights of the BLSTM network. The procedure were applied to all cases in order to establish the base system for comparison of the proposals.
- *Case 2 (Transfer from AAN):* In this case, an auto-associative network (AAN) is trained from clean speech parameters. AAN approximates the identity function by presenting the same information at both the inputs and outputs. The set of internal weights resulting from this training are transferred to all the BLSTM network to begin its adjustment.

- *Case 3 (Transfer from White Noise):* In this approach, a BLSTM network is trained for denoising the signal with SNR0 White Noise, and then the results are transferred to the whole set of BLSTM networks, both for denoising White and Babble Noise at all SNR levels.
- *Case 4 (Transfer from Babble Noise):* Similar to the previous case, but the first BLSTM autoencoder is trained for denoising SNR0 of Babble Noise, and then the results are transferred to the whole set of BLSTM networks, both for denoising White and Babble Noise at all SNR levels.

The experimental study aims to numerically compare and recommend the best procedure in terms of less training time, whilst the capacity of the network in denoising the speech increases or is not affected considerable, according to the objective quality evaluation measures.

### Evaluation

The following common measures for artificial neural network training were applied to the results given by the four cases of initialization and transfer learning:

- *SSE (sum of squared errors):* Is the sum of the squared differences between each output and the expected value in the validation set, calculated at each training epoch. For a given network θ, SSE is computed as:

  $$SSE(\theta) = \sum_{n=1}^{T} (c_x - \hat{c}_x)^2, \quad (4)$$

  where $c_x$ is the real value of the parameters in the validation set (*x(t)* in Equation (1)), $\hat{c}_x$ is the predicted output from each frame of the validation set ($\hat{\mathbf{x}}(t)$ in Equation (3)), and *T* the total number of frames from each audio file.
- *Number of epochs:* An epoch is a complete cycle of the parameters in the dataset presented to network in order to adjust the update the internal weights of the BLSTM networks. The total time taken to train any neural network is the sum of the time taken by individual epochs required in the process.

All the BLSTM networks were trained on a Linux computer with Pentium i7 Processor accelerated with an Nvidia GPU.

## Results

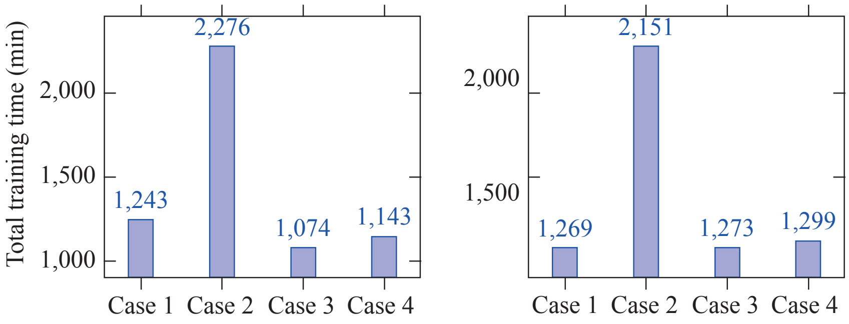
In table 1, the results of each of the four cases are presented. The missing value in the Transfer-White results is because this particular SNR level was used to initialize the rest of the BLSTM autoencoders, so there no information transferred from any other network to this SNR level.

**Table 1.** SSE results and number of epochs required for training (in parenthesis) for each case of White Noise analysed. * is the best result for each SNR level.

| Case 1 | | | Case 2 | Case 3 | Case 4 |
|---|---|---|---|---|---|
| | | | SNR-10 | | |
| Random 1 | Random 2 | Random 3 | Transfer-AAM | Transfer-White* | Transfer-Babble |
| 464.45(271) | 472.70(159) | 458.45(205) | 463.06(196) | 447.94(96) | 457.52(211) |
| | | | SNR-5 | | |
| Random 1 | Random 2 | Random 3 | Transfer-AAM | Transfer-White* | Transfer-Babble |
| 401.41(168) | 396.32(180) | 395.52(188) | 392.07(207) | 380.63(162) | 392.17(168) |
| | | | SNR0 | | |
| Random 1 | Random 2 | Random 3 | Transfer-AAM* | Transfer-White | Transfer-Babble |
| 332.67(196) | 328.65(269) | 332.64(262) | 328.28(374) | - | 330.99(194) |
| | | | SNR5 | | |
| Random 1 | Random 2 | Random 3 | Transfer-AAM* | Transfer-White | Transfer-Babble |
| 239.98(198) | 283.25(362) | 290.52(221) | 283.22(394) | 285.40(157) | 288.82(275) |
| | | | SNR10 | | |
| Random 1 | Random 2 | Random 3 | Transfer-AAM* | Transfer-White | Transfer-Babble |
| 251.86(376) | 251.80(340) | 253.68(333) | 243.79(655) | 245.95(417) | 253.61(295) |

According to the results presented in Table 1, in SNR-10 level, transferring the set internal weights from white noise and babble noise showed better results than from auto-associative memory and random initialization. Additionally, they obtained the result in less time and this is reflected in the number of epochs (case 1). Similarly, in SNR-5 the best result was obtained using Transfer-White: the smallest SSE value and training time (case 3).

As to the noise levels from SNR0 and on, the best results were obtained using AAM. However, once again the results obtained with transfer (from both whiteand babble noise) outperform the results obtained with random initialization. Additionally, in the transfer cases, the total training time of the networks is reduced, giving these an advantage in computational cost.



(a) White Noise (b) Babble Noise

**Figure 1.** Total training time for denoising BLSTM networks

Transfer using AAM requires an additional training process that is not used for the denoising task but just for initialization purposes. Due to this additional training process, the total training time grows. In general, the good results obtained by the AAM came as a trade-off with training time.

The Figure 1a shows that the biggest average training time of the networks is case 2: 2151 min. This training time is almost 83.11% more in comparison to random initialization (case 1). In contrast, in cases 3 and 4, the results were better than case 1 and are competitive with case 2 in some levels, requiring a notably short training time between 47.19%-50.22% less than case 2.

In Table 2, the results of each of the four cases are presented. The missing value in the Transfer-Babble results is because this particular SNR level was used to initialize the rest of the BLSTM networks, and the best result was obtained transferring from auto-associative memory for four of five levels of noise; however, the computational cost is higher when compared with other cases. This is shown graphically in Figure 1b.

An advantage of using non-random procedures is that it is no longer necessary to use multiple trials to obtain significant good results compared to others, because of the great variability present in random initializations. For example, in Table II, for SNR-10 there are ranges from 585.85 to 603.75 for SSE values in Case

1. Instead, with the transfer procedures, the results obtained are competitive and (in some cases even better) with lower training time.

## Conclusions

In this work, an experimental approach was performed to compare the efficiency of training denoising BLSTM networks, for natural and artificial noises. The results allow the numerical validation of the transfer learning procedure for regression problems in this kind of network, to establish that the random and independent initialization of a set of BLSTM networks is not the best option in terms of efficiency and training time. Transferring the weights from one BLSTM trained with a particular condition (such an SNR level) to initialize the weights of the rest of the conditions and noise types can reduce training time and increase efficiency.

**Table 2.** SSE results and number of epochs required for training (in parenthesis) for each case of Babble Noise analysed. * is the best result for each SNR level.

| Case 1 | | | Case 2 | Case 3 | Case 4 |
|---|---|---|---|---|---|
| | | | SNR-10 | | |
| Random 1 | Random 2 | Random 3* | Transfer-AAM | Transfer-White | Transfer-Babble |
| 603.74(127) | 595.22(129) | 585.85(139) | 586.36(163) | 593.13(102) | 595.17(67) |
| | | | SNR-5 | | |
| Random 1 | Random 2 | Random 3 | Transfer-AAM* | Transfer-White | Transfer-Babble |
| 491.71(122) | 488.73(165) | 494.40(129) | 466.43(185) | 499.29(79) | 481.43(47) |
| | | | SNR0 | | |
| Random 1 | Random 2 | Random 3 | Transfer-AAM* | Transfer-White | Transfer-Babble |
| 374.24(175) | 370.51(164) | 372.14(189) | 356.54(234) | 369.20(220) | - |
| | | | SNR5 | | |
| Random 1 | Random 2 | Random 3 | Transfer-AAM* | Transfer-White | Transfer-Babble |
| 289.60(245) | 282.32(279) | 285.17(302) | 266.12(587) | 283.73(317) | 280.34(400) |
| | | | SNR10 | | |
| Random 1 | Random 2 | Random 3 | Transfer-AAM* | Transfer-White | Transfer-Babble |
| 221.24(636) | 230.20(484) | 225.41(522) | 219.36(590) | 220.42(587) | 221.62(609) |

In particular, the transfer of information from the BLSTM trained with white noise present the best results. It can be explained for the most homogeneous distribution of values that represent the white noise, but further experimentation should be conducted in order to validate this hypothesis. The transfer of information from Auto-associative Memories presents the best results in most cases, but the time required for its training is considerably higher.

It can be stated that random initialization for the complete set of neural networks is in no case the best option. Transfer from a particular network is the best option in terms of time and results, which can be the best option in exploratory studies of regression, such as exploring architectures or comparing neural networks. The transfer from Auto-associative Memories can be considered only to achieve the best results when the training time is not an issue.

For future work, more extensive validation of the transfer learning among some types of noise or particular SNR levels can be performed. Statistical validation of the improvements achieved could be relevant, along with numerical validation of the results in terms of the quality of the signal.

## References

[1] Weninger, F., Watanabe, S., Tachioka, Y., and Schuller, B. “Deep recurrent denoising auto-encoder and blind de-reverberation for reverberated speech recognition.” IEEE ICASSP, 2014.

[2] Donahue, Chris, Bo Li, and Rohit Prabhavalkar. “Exploring speech enhancement with generative adversarial networks for robust speech recognition.” IEEE ICASSP, 2018.

[3] Coto-Jiménez, Marvin, John Goddard-Close, and Fabiola Martínez-Licona. “Improving automatic speech recognition containing additive noise using deep denoising autoencoders of LSTM networks.” International Conference on Speech and Computer. Springer, Cham, 2016.

[4] Abouzid, Houda, et al. “Signal speech reconstruction and noise removal using convolutional denoising audio-encoders with neural deep learning.” Analog Integrated Circuits and Signal Processing 100.3 (2019): 501-512.

[5] Ling, Zhang. ”An Acoustic Model for English Speech Recognition Based on Deep Learning.” 2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA). IEEE, 2019.

[6] Coto-Jiménez, M.; Goddard-Close, J.; Di Persia, L.; Rufiner, H.L. “Hybrid Speech Enhancement with Wiener filters and Deep LSTM Denoising Autoencoders.” In Proceedings of the 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), San Carlos, CA, USA, 18–20 July 2018; pp. 1–8.

[7] González-Salazar, Astryd, Michelle Gutiérrez-Muñoz, and Marvin Coto-Jiménez. ”Enhancing Speech Recorded from a Wearable Sensor Using a Collection of Autoencoders.” Latin American High Performance Computing Conference. Springer, Cham, 2019.

[8] Gutiérrez-Muñoz, Michelle, Astryd González-Salazar, and Marvin Coto-Jiménez. “Evaluation of Mixed Deep Neural Networks for Reverberant Speech Enhancement.” Biomimetics 5.1 (2020): 1

[9] Tkachenko, Maxim, et al. “Speech Enhancement for Speaker Recognition Using Deep Recurrent Neural Networks.” International Conference on Speech and Computer. Springer, Cham, 2017.

[10] Liu, Ming, et al. “Speech Enhancement Method Based On LSTM Neural Network for Speech Recognition.” 2018 14th IEEE International Conference on Signal Processing (ICSP). IEEE, 2018.

[11] Weiss, Karl, Taghi M. Khoshgoftaar, and DingDing Wang. “A survey of transfer learning.” Journal of Big Data 3.1 (2016): 9.

[12] Song, Guangxiao, et al. “Transfer Learning for Music Genre Classification.” International Conference on Intelligence Science. Springer, Cham, 2017.

[13] Yeom-Song, Víctor, Marisol Zeledón-Córdoba, and Marvin Coto-Jiménez. ”A Performance Evaluation of Several Artificial Neural Networks for Mapping Speech Spectrum Parameters.” Latin American High Performance Computing Conference. Springer, Cham, 2019.

[14] Erro, Daniel, et al. “HNM-based MFCC+ F0 extractor applied to statistical speech synthesis.” IEEE ICASSP, 2011.

[15] Greff, Klaus, et al. “LSTM: A search space odyssey.” IEEE transactions on neural networks and learning systems 28.10 (2016): 2222–2232.

[16] Kominek, John, and Alan W. Black. “The CMU Arctic speech databases.” Fifth ISCA workshop on speech synthesis. 2004.