Hybrid Speech Enhancement with Wiener filters and Deep LSTM Denoising Autoencoders

1st Marvin Coto-Jiménez
Pattern Recognition and Intelligent Systems Lab (PRIS-Lab)
Escuela de Ingeniería Eléctrica, Universidad de Costa Rica
San José, Costa Rica
marvin.coto@ucr.ac.cr

2nd John Goddard-Close
Departamento de Ingeniería Eléctrica
Universidad Autónoma Metropolitana
Mexico City, Mexico
jcg@xanum.uam.mx

3rd Leandro Di Persia
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i)
FICH-UNL-CONICET
Santa Fe, Argentina
ldipersia@sinc.unl.edu.ar

4th Hugo Leonardo Rufiner
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i) & Laboratorio de Cibernética
FICH-UNL-CONICET & Facultad de Ingeniería, UNER, Oro Verde
Entre Ríos, Argentina
lrufiner@ingenieria.uner.edu.ar

***Abstract*—Over the past several decades, numerous speech enhancement techniques have been proposed to improve the performance of modern communication devices in noisy environments. Among them, there is a large range of classical algorithms (e.g. spectral subtraction, Wiener filtering and Bayesian-based enhancement), and more recently several deep neural network-based. In this paper, we propose a hybrid approach to speech enhancement which combines two stages: In the first stage, the well-known Wiener filter performs the task of enhancing noisy speech. In the second stage, a refinement is performed using a new multi-stream approach, which involves a collection of denoising autoencoders and auto-associative memories based on Long Short-term Memory (LSTM) networks.**

We carry out a comparative performance analysis using two objective measures, using artificial noise added at different signal-to-noise levels. Results show that this hybrid system improves the signal's enhancement significantly in comparison to the Wiener filtering and the LSTM networks separately.

***Index Terms*—Deep learning, Denoising autoencoders, LSTM, Signal processing.**

I. Introduction

The enhancement of speech in the presence of noise has been a topic of interest over the past several decades, giving that speech signals are often adversely affected in real world environments through the introduction of different types of noise and reverberation. Communication devices and systems may be affected in their quality and recognition performance [1]–[4] with such noise addition to the speech information.

A speech enhancement algorithm can be viewed as successful if it suppresses perceivable background noise, and preserves or enhances perceived signal quality [5].

For the task of improving speech recognition systems and enhance speech signals, deep neural networks (DNN) have been presented in [6]–[9]. One approach that has been applied successfully is that of mapping spectral features from noisy speech into the features of the corresponding clean speech, using autoencoders based on perceptrons or recurrent neural networks (RNNs).

Among the new types of RNNs, the Long Short-Term Memory Network (LSTM) has succeeded in mapping noisy or reverberant speech parameters to clean speech, by using features derived from the spectrum, usually MFCC. These features are of interest and have been used widely because automatic speech recognition systems are frequently based on them. One recent line of research is to include additional information in this approach e.g. fundamental frequency (f_0) and energy coefficients, to improve the obtained results.

After considering both traditional signal processing-based methods, such as the Wiener filtering, and the deep neural networks approach, we are proposing a hybrid denoising method in two stages. Benefits from this type of speech enhancement can be applied to mobile phones, VoIP, speech recognition, and are especially important for hearing-impaired listeners, given their particular difficulty in noisy backgrounds [10].

A. Related work

Some techniques for feature enhancement of speech signals based on deep learning have been presented recently. These techniques rely on the enhancement of features derived from the spectrum, typically MFCC. For example, MFCCs (cepstral coefficients plus its first and second delta) are used in [11],

[12], and super-vectors, with 24 MFCCs formed by splicing together 9-frame windows of MFCCs in [13].

Some of these techniques have outperformed other denoising algorithms on speech recognition tasks, where the speech signals contain noise of different types with various signal-to-noise levels [14]–[16]. It has been observed the advantage in reducing the annoying musical artifact commonly present in classical speech enhancement algorithms [17].

The principal mechanism for enhancing speech signals using deep learning algorithms is to use deep neural networks as regression models which map the noisy speech into the corresponding clean speech [1] [2]. This approach has been successfully applied to speech signals obtained from speakers under different conditions, such as noise types and signal-to-noise ratios and single and two channel comparison.

Reverberant speech has also been analyzed using a similar framework in [18], [19], by mapping reverberant MFCC to clean ones. These mappings have been used for speech recognition, for example, in [20], and include a hybrid deep neural network/HMM approach. Also, the enhancement of speech signals with background music has been successfully tested with deep neural networks [21].

Features other than MFCC, such as 13-dimensional Perceptual Linear Prediction (PLP) with windowed mean-variance normalization and up to third-order derivatives, have also been tested [22].

LSTM networks for speech enhancement have been presented previously in [23], using also MFCC as features, and a single step of networks in the enhancement. In the present paper a two steps approach is presented, combining Wiener filters and LSTM networks.

The use of several steps for the speech enhancement problem has been approached stacking many denoising autoencoders in [24]. Deep learning algorithms have been employed as estimators of speech enhancement techniques such as Wiener filters in [3]. To our knowledge, the opposite direction remains unexplored before our proposal: training deep network-based algorithms using Wiener filters or other techniques as its inputs.

B. Overview

In this paper, we present a new approach for enhancing noisy speech, by considering a hybrid two stages of Wiener filtering and a collection of LSTM networks that maps the output of the Wiener filter to clean features. The whole system is trained and tested with examples from the Carnegie Mellon University speech database. Several objective measures are used to test the results, which show the benefits of our proposed method.

The rest of this paper is organized: Section II gives the background and context of the problem of denoising, Section III presents the proposed hybrid systems. Section IV describes the experimental setup, Section V presents the results with a discussion, and finally, in Section VI, we present the conclusions.

II. Background

A. Problem Statement of Speech Enhancement

In speech enhancement, a speech signal degraded with additive noise is processed so as to improve its quality with respect to factors such as intelligibility or perceptual quality.

We can assume that the corrupted signal, y, is the sum of a speech signal, x, and noise d, given by:

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

Applying the Short-time Fourier Transform, in the spectral domain, the formulation of the problem becomes:

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

where k is the frequency index and n the time-segment index. In classical methods, $x(t)$ is considered uncorrelated to $d(t)$, and a broad class of speech enhancement algorithms estimate $X_k(n)$ from the power spectral domain of $x(t)$ and $d(t)$. Among them, the Wiener filter, which was presented for the first time in the decade of 1940, aims is to filter a noise-corrupted input, while output an estimate of the original signal, based on statistical computations of the noise. Let $\tilde{x}(t)$ be the estimation of the signal using this algorithm.

In deep learning approaches, $x(t)$ (or $X_k(n)$) can be estimated using algorithms that learn an approximated function $f(\cdot)$ between the noisy and clean data of the form:

$$\hat{x}(t) = f\left(y(t)\right). \tag{3}$$

The precision of the approximation $f(\cdot)$ usually depends on the amount of training data and the algorithm selected.

In a hybrid approach combining both systems, the $\hat{x}$ of the deep learning algorithm is approximated from the output of the Wiener filter, so the relation can be established as

$$\hat{\tilde{x}}(t) = f\left(\tilde{x}(t)\right). \tag{4}$$

It is expected that $\hat{\tilde{x}}(t)$ is a better estimation of $x(t)$ than $\tilde{x}(t)$ and $\hat{x}(t)$. The idea is that the Wiener filter can reduce the noise, but at the cost of introducing artifact. Then the deep LSTM network would eliminate the artifact, preserving the good information produced by the Wiener filter.

B. Wiener filter

Speech enhancement techniques commonly estimate a short-term suppression factor, adjusted for each frequency component with *a posteriori* signal-to-noise ratio. Some recent techniques have included an *a priori* signal-to-noise ratio for the computation of the adjustment in the enhancement process [25]. These factors can be estimated using several techniques, for example Wiener estimation.

We are using an implementation of this approximation that defines the Noise Power Spectral Density $\hat{P}^t(\cdot)$ at each frequency component f_k as:

$$\hat{P}_B^t(f_k) = \lambda \hat{P}_B^{t-1}(f_k) + (1-\lambda)|B^t(f_k)|^2 \tag{5}$$

where B is the spectrum of the noise, $P(\cdot)$ denotes half-wave rectification at the time interval t.

The proposed *a priori* signal-to-noise ratio is defined as:

$$S\hat{N}R^{t}_{prio}(f_k) = (1-\beta)P[S\hat{N}R^{t}_{prio}(f_k) - 1] + \beta\frac{|\hat{S}^{t-1}(f_k)|^2}{\hat{P}_B(f_k)} \quad (6)$$

Further details can be found in [25]. In our implementation, we fix the main parameters to the values $\lambda = \beta = 0.98$.

C. Long Short-term Memory Neural Networks

Several kinds of neural networks have been tested for classification and regression purposes over the past several decades. Recently, new kinds of networks organized in many layers achieved good results in many problems of a wide spectrum of applications. From the emergence of RNNs for modeling sequential parameters, e.g. in human speech, handwriting recognition and synthesis [26] [27], new trends in modeling the dependent nature of sequential information have been opened. RNNs can store information by feedback connections between neurons in the hidden layers to themselves or others neurons in the same layer.

With the aim of storing information in the short and the long term, LSTM networks presented in [28] have introduced a set of gates and memory cells that control the access, writing and propagation of memory values over the network. LSTM networks presented encouraging results in speech recognition and music composition, which heavily depends on previous states of the information [29] [30] [28].

Figure 1 illustrates a unit (substituing the neuron) in the hidden layer of the network. The four gates controls the operations of input, output, and erasing (forget gate) the memory value. More details on the training procedure and the mathematical modeling of the LSTM can be found in [31].

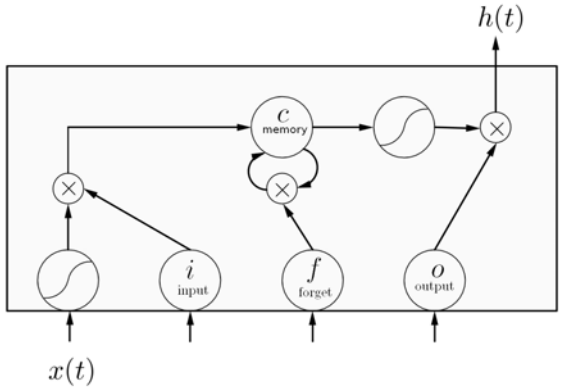


Fig. 1. Unit memory of a LSTM. $x(t)$ is the input and $h(t)$ the output of the unit. i, o, f represent the gates, and c the value of the memory.

D. Denoising with Deep Neural Networks

The idea of training a neural networks in speech enhancement and denoising was first introduced about thirty years ago, with networks trained and tested on binary input patterns, corrupted by randomly flipping a fraction of the input bits.

Other than binary inputs, acoustic coefficients were modeled with a single layer a few years later. Neither the computer capabilities nor the algorithms were adequate for including more hidden layers or considering much larger sets of data, so the benefits could not encourage many more experiences [16].

As usual, parameters of the DNN are found using training data in order to minimize the average reconstruction of the input, that is, to have output $f(y)$ as close as possible to the uncorrupted signal x [32].

One of the recent architectures of neural networks that have achieved considerable success is called a denoising autoencoder, consisting of two steps: the first one is the encoder, which performs a mapping f that transforms an input vector y into a representation h in the hidden layers. The second step is the decoder, which mapped back the hidden representation into a vector $\hat{x}$ in input space.

For this purpose, during the training stage, noise corrupted features are presented at the inputs of the autoencoders, while the corresponding clean features of the same dimensionality became the outputs. The training algorithm adjusts the parameters of the network in order to learn the complex relationships between them.

III. Proposed system

In order to improve the enhancement of noisy utterances, we apply as a first stage the Wiener filter at the noisy utterances. In the second stage, we parametrize the waveform at the output of the filter and train a autoencoders of LSTM. Each network maps the Wiener-enhanced parameters $\tilde{x}$ to clean parameters x.

Figure 2 illustrate this procedure for denoising 39 MFCC, as used in this paper. After training, for an input vector y, the network produces an output $\hat{x}(y, \mathcal{W})$ which depends on the input y and the parameters $\mathcal{W}$ of the network (i.e., the set of weights). The purpose of training is to ensure that the outputs represent a closer version of the correspondent clean vector x. The process can be expressed in terms of the objective function [33]:

$$\min_{\mathcal{W}} \mathbb{E}\left[x - \hat{x}(y, \mathcal{W})\right]. \quad (7)$$

Also, the process requires the training of one autoencoder for each noise type and level.

IV. Experimental setup

We shall describe in some detail the experimental setup that was followed in the paper. The whole process, from data generation to evaluation, can be summarized in the following steps:

1) Noisy database generation: Files containing different types of noise were generated and added to each audio file in the database for a given signal-to-noise ratio (SNR). Two types of noise were generated and added: White Noise and Pink Noise. Five noise levels were

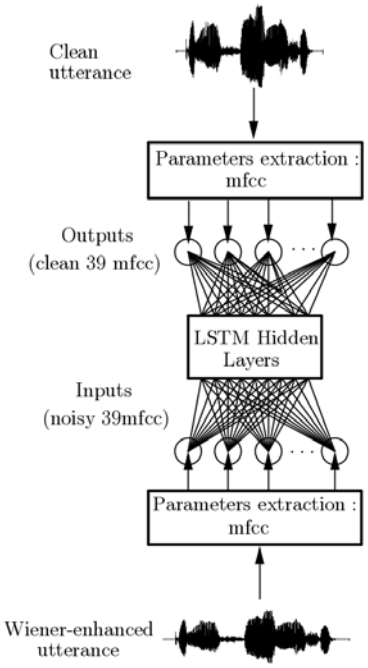


Fig. 2. Denoising autoencoder with LSTM units in hidden layers

added, in order to cover a range from light to heavy noise levels for each noise type.

2) Feature extraction and input-output correspondence: A set of parameters was extracted from the noisy, the Wiener-enhanced and the clean audio files. Those from the noisy files (or from the Wiener-filtered) were used as inputs to the networks, while the corresponding clean features were the outputs.
3) Training: During training, using forward pass and back-propagation through time algorithm, the weights of the networks were adjusted as the noisy and clean utterances were presented at the inputs and at the outputs. A total of 900 utterances (about 80% of the total database) were used for training. Details and equations of the algorithm followed can be found in [34].
4) Validation: After each training step, the sum of squared errors were computed within the validation set of 182 utterances (about 15% of the total database), and the weights of the network updated in each improvement.
5) Test: A subset of 50 randomly selected utterances (about 5% of the total amount of utterances of the database) was chosen for the test set. These utterances were not part of the training process, to provide independence between the training and testing. The same 50 sentences were also used for the results obtained with the Wiener filter alone and the LSTM networks without the hybrid system.

In the following subsections, further details of the main experimental setup are given.

A. Database

In our work, we chose the SLT voice from the CMU ARCTIC databases [35], designed for speech research. The whole set of 1132 sentences were used to randomly define the training, validation and test sets. In our work, we chose the female SLT voice, and the whole set of 1132 sentences were used to randomly define the training (849 sentences), validation (233 sentences) and test sets (50 sentences).

B. Noise

Artificially generated White and Pink noise was added to the waveforms of the database to achieve a desired SNR. The SNR levels considered for the experiments were selected according to the criteria of degrading the speech signal from heavy to light in the three cases of added noises.

C. Feature extraction

The audio files of the noisy and the Wiener-filtered database were downsampled to 16kHz, 16 bits, to extract parameters using the Ahocoder system [36]. A frame size of 160 samples and frame shift of 80 samples were used to extract 39 mfcc, f_0 and energy of each sentence.

After the enhancement, a waveform can be synthesized from parameters using the same Ahocoder system, from a sequence of 41-dimensional vectors, using separate files for f_0 and energy $+$ mfcc (enhanced) values.

D. Evaluation

Two objective measures were selected to evaluate the results given by the different enhancement methods:

- PESQ: This measure uses a psychoacoustic model to predict the subjective quality of speech. This measure is defined in the ITU-T recommendation P.862.ITU. Results are given in interval $[0.5, 4.5]$, where 4.5 corresponds to a perfect reconstruction of the signal [37].
- Weighted-slope spectral distance (WSS): This measure calculates the weighted difference between spectral slopes in frequency bands, by measuring the difference between adjacent spectral magnitudes [38].

E. Experiments

For comparison purposes, we kept the output of the Wiener filter, and also trained another set of LSTM networks to directly map the noisy features to clean features. The base system is described following the nomenclature:

- LSTMA: One autoencoder LSTM network is used to enhance 39 MFCCs, leaving the noisy energy and f_0 parameter without enhancing.
  Each LSTM (one for each noise level) was trained to map the corresponding noisy features (or enhanced by the Wiener filter) of the waveform to clean features. The proposed hybrid system is described using the following nomenclature:
- HW-LSTMA: On the first stage, the Wiener filter is applied to the waveform. At the output of the Wiener filter, one denoising LSTM is used to enhance 39 MFCCs,

leaving the energy and f_0 parameter of the Wiener filter output.

Each sentence was parameterized using the Ahocoder system. The LSTM architecture for the networks was defined by trial and error. Initially, we considered a single hidden layer with 50 units and then increased the size with steps of 50 units, up to three hidden layers with 300 units in each layer. The final selection consisted of a network with three layers containing 150, 100 and 150 units in each one.

This network gave the best results in the trial experiments, and also had a manageable training time, considering that we use 20 LSTM networks in this work. The training procedure was accelerated by a NVIDIA GPU system, taking about 7 hours to train each LSTM.

V. Results and Discussion

The results are organized into two parts. In the first part (Section V.A), we present the average scores of objective measurements of the test set for the hybrid system and its comparison to the Wiener filtering and LSTM system alone. In the second part V.B, statistical test are applied to determine whether or not each system represent statistically significant enhancement of the noisy signal.

A. Objective measures

The results for WSS for the two kinds of noise and the five noise levels are presented in Table I. The hybrid system HW-LSTMA achieves the best results for three levels of white noise, corresponding to the higher levels. The most remarkable improvement for the White noise with the Hybrid system was at SNR-10, where HW-LSTMA obtained a 45% better value than the Wiener filter. The LSTM autoencoder achieves best results in the lighter levels of noise.

For the Pink noise, the hybrid system achieves best result for the higher SNR, where the LSTMA obtained best results in the rest of levels. The hybrid approach performs similar in most of cases, and both succeed in enhancing significantly the noisy signal, as described in the next subsection.

The results for PESQ are shown in Table II. For this objective measure, the hybrid HW-LSTMA gives the best results in every level of White Noise. For SNR-10, the PESQ raised 50% with the Hybrid system compared to the Wiener Filter. In SNR-5 the PESQ was 54% better with the Hybrid system than the Wiener filter, and 82% better than the LSTM alone.

For the case of Pink noise, the Wiener filter alone achieved the best result for the higher noise level, but in the rest of levels, the hybrid HW-LSTMA present the best results, sharing the best result in some noise levels with the LSTMA. For the ten cases of noise types and levels, the hybrid system achieved the best PESQ in nine cases.

Figure 3 illustrate some spectrograms of the same utterance, from the noisy to the clean version filtered with Wiener and the Hybrid proposal.

B. Statistically significant enhancement of the noisy speech signal

In this section, we present a statistical analysis in order to determine when the results presented so far significantly enhance the noisy speech signal. One reason for this is the fact that an algorithm may give the best result for a measure without significantly enhancing the noisy signal.

For the statistical analysis, we applied Tukey's HSD test [39] [40] to assess significant differences between the enhanced speech signal and the noisy signal. This test gives pairwise comparisons between all results and the Tables III-IV report which of the algorithms significantly improve the noisy speech utterances.

In Table III, we see that hybrid HW-LSTMA achieves significant enhancement for WSS at all noise levels of White noise, succeeding among the rest of systems. For Pink noise, the results of HW-LSTMA and LSTMA also improved the noisy signal significantly

In the case of PESQ, Table IV, the results for White noise are similar concerning the enhancing the noisy signal at all noise levels for the hybrid system. Wiener and the LSTMA's improved significantly only at the lower noise levels. For Pink noise, the results of significance are similar for all the algorithms, where it can also be noticed that none of them did improve the noisy signal for this measure at the higher noise level.

From these results, it can be established that at high SNR levels, it is convenient to use only the LSTM networks, which have a better capacity to recover information from the noisy signal. This can be explained given that the frequency information has not been corrupted significantly. At these levels, with the hybrid approach, some information is lost and the process does not result in benefits for the purpose of enhancing the signal. This happens for more values of SNR in White noise than in Pink noise.

When noise is presented at lower SNR levels (SNR -5, SNR -10), the useful frequency information is hardly detectable on the noisy signal. Here, the hybrid approach that first uses the Wiener filter presents evident benefits, by eliminating part of the noise component and allowing the LSTM networks to reconstruct the content in a better way, as shown previously in Figure 2.

VI. Conclusions

In this work, we have presented a new proposal for speech enhancement using a hybrid approach consisting on Wiener filters and deep neural networks.

We conducted an extensive comparison with the single Wiener filter and the LSTM networks alone, and performed statistical tests to assess the statistical significance in enhancing the noisy speech signals.

We evaluated the proposals using speech utterances taken from a well-known speech database. Two different types of noise, White and Pink were added to the utterances at five SNR levels. The evaluations were performed using two common objective quality measures.

TABLE I

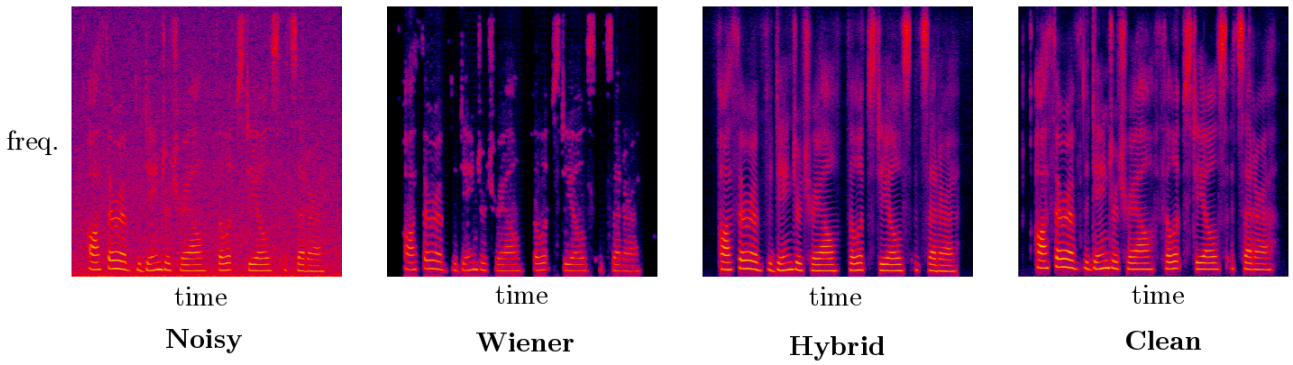
WSS RESULTS (AVERAGE MEASURES FOR THE TEST SET). THE LOWER VALUES REPRESENT BETTER RESULTS. * IS THE BEST RESULT.

White Noise Enhancement	SNR -10	SNR -5	SNR 0	SNR 5	SNR 10
None	49.7	42.6	37.1	32.9	30.0
Wiener	68.8	57.8	47.8	40.2	34.5
LSTMA	72.0	54.9	27.8	21.0*	17.0*
HW-LSTMA	38.5*	30.6*	26.6*	23.7	20.0
Pink Noise Enhancement	SNR -10	SNR -5	SNR 0	SNR 5	SNR 10
None	63.6	55.1	47.0	39.4	33.3
Wiener	79.2	67.5	33.7	41.9	34.1
LSTMA	70.9	41.3*	28.3*	19.7*	15.2*
HW-LSTMA	64.8*	45.0	29.5	22.8	18.1

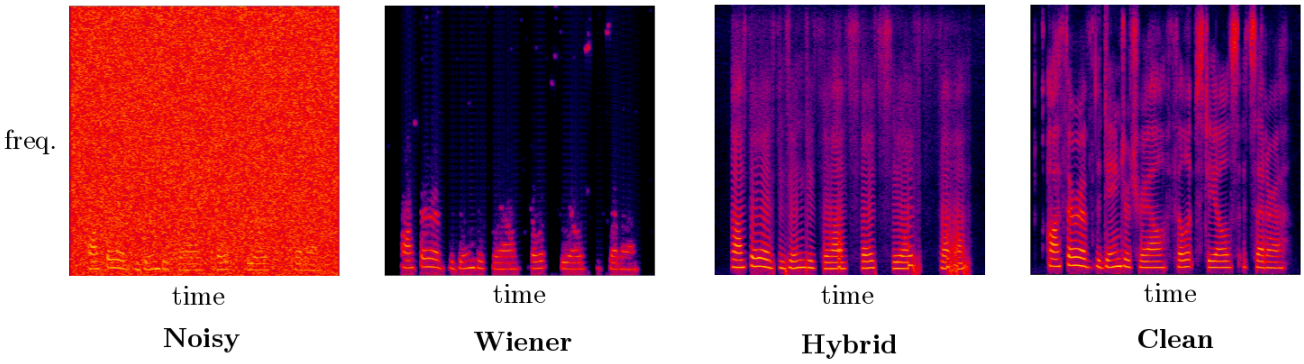
TABLE II

PESQ RESULTS (AVERAGE MEASURES FOR THE TEST SET). THE HIGHER VALUES REPRESENT BETTER RESULTS. * IS THE BEST RESULT.

White Noise Enhancement	SNR -10	SNR -5	SNR 0	SNR 5	SNR 10
None	1.0	1.1	1.3	1.6	1.9
Wiener	1.0	1.3	1.7	2.1	2.4
LSTMA	0.8	1.1	1.8	2.4*	2.7*
HW-LSTMA	1.5*	2.0*	2.2*	2.4*	2.7*
Pink Noise Enhancement	SNR -10	SNR -5	SNR 0	SNR 5	SNR 10
None	0.8	1.1	1.3	1.7	2.0
Wiener	1.0*	1.3*	1.9	2.1	2.5
LSTMA	0.7	1.3*	1.9	2.5*	2.9*
HW-LSTMA	0.7	1.3*	2.1*	2.5*	2.9*



(a) Pink Noise results for SNR 10



(b) White Noise results for SNR -10

Fig. 3. Spectrograms for the noisy, clean and enhanced utterances

TABLE III
WSS RESULTS. TICKS INDICATE SIGNIFICANT ENHANCEMENT OF NOISY SPEECH

White Noise Enhancement	SNR -10	SNR -5	SNR 0	SNR 5	SNR 10
Wiener					
LSTMA			✓	✓	✓
HW-LSTMA	✓	✓	✓	✓	✓
Pink Noise Enhancement	SNR -10	SNR -5	SNR 0	SNR 5	SNR 10
Wiener			✓		
LSTMA		✓	✓	✓	✓
HW-LSTMA		✓	✓	✓	✓

TABLE IV
PESQ RESULTS. TICKS INDICATE SIGNIFICANT ENHANCEMENT OF NOISY SPEECH

White Noise Enhancement	SNR -10	SNR -5	SNR 0	SNR 5	SNR 10
Wiener			✓	✓	✓
LSTMA			✓	✓	✓
HW-LSTMA	✓	✓	✓	✓	✓
Pink Noise Enhancement	SNR -10	SNR -5	SNR 0	SNR 5	SNR 10
Wiener		✓	✓	✓	✓
LSTMA		✓	✓	✓	✓
HW-LSTMA		✓	✓	✓	✓

The results show that our hybrid approach achieves better results than the Wiener filter and the LSMT networks alone in enhancing the speech signals for the majority of noise levels and noise types.

The main shortcoming of out hybrid approach is the computational cost of training the number of networks required, because kind of noise and every noise level requires independent networks, and each one a training time of about seven hours. After training, applying the LSTM networks for noise reduction in test sentences requires less than a second to enhance an utterance.

The computational cost is an obstacle, for example, to the time needed for finding good network architectures and suitable training parameters. However, the results have shown the capacity of the hybrid systems to significantly enhance noisy speech of different types. For implementation in devices or applications, additional noise aware systems should be integrated to match the specific noise type and level of the trained networks.

Future work will include the new combination of hybrid algorithms, to improve the enhancing of the noisy speech, as well as exploring ways of reducing the computational cost of training. Finally, new types of noise and multiple-noise conditions could be considered for the proposal, with additional evaluation measurements.

ACKNOWLEDGMENTS

This work was supported by the Universidad de Costa Rica (in Costa Rica), SEP and CONACyT under the Program SEP-CONACyT, CB-2012-01, No.182432, in Mexico; also by *Universidad Nacional de Litoral* (with PACT 2011 #58, CAI+D 2011 #58-511) and *Consejo Nacional de Investigaciones Científicas y Técnicas* (CONICET) from Argentina.

REFERENCES

[1] F. Weninger, S. Watanabe, Y. Tachioka, and B. Schuller, "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition." In 2014 IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4623-4627.

[2] F. Weninger, et al. "Feature enhancement by deep LSTM networks for ASR in reverberant multisource environments." Computer Speech & Language 28.4 (2014): 888-902.

[3] A. Narayanan, and W. DeLiang. "Ideal ratio mask estimation using deep neural networks for robust speech recognition." In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[4] D. Bagchi, et al. "Combining spectral feature mapping and multi-channel model-based source separation for noise-robust automatic speech recognition". In Proceedings of the IEEE ASRU.

[5] J. Hansen, and B.L. Pellom. "An effective quality evaluation protocol for speech enhancement algorithms." Proceedings of the ICSLP. Vol. 7. 1998.

[6] J. Du et al. "Robust speech recognition with speech enhanced deep neural networks." In Proceedings INTERSPEECH 2014, pp. 616-620.

[7] K. Han, et al. "Deep neural network based spectral feature mapping for robust speech recognition". In Proceedings of INTERSPEECH 2015, pp. 2484-2488.

[8] A.L. Maas, et al. "Recurrent Neural Networks for Noise Reduction in Robust ASR." In Proceedings of INTERSPEECH 2012, pp. 22-25.

[9] L. Deng, et al. "Recent advances in deep learning for speech research at Microsoft." In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 8604-8608.

[10] E.W. Healy, et al. "An algorithm to improve speech recognition in noise for hearing-impaired listeners." The Journal of the Acoustical Society of America 2013, 134.4, pp. 3029-3038.

[11] M. Seltzer, Y. Dong and Y. Wang. "An investigation of deep neural networks for noise robust speech recognition." In IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 7398-7402.

[12] O. Abdel-Hamid,et al. "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition." IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 4277-4280.

[13] J. Huang, and B. Kingsbury. "Audio-visual deep learning for noise robust speech recognition." IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 7596-7599.

[14] A.L. Maas, et al. "Recurrent Neural Networks for Noise Reduction in Robust ASR." Proceedings INTERSPEECH 2012.

[15] A. Kumar and F. Dinei. “Speech Enhancement In Multiple-Noise Conditions using Deep Neural Networks.” arXiv preprint arXiv:1605.02427 (2016).
[16] G. Hinton, et al. “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups.” IEEE Signal Processing Magazine, 2012, 29.6, pp. 82-97.
[17] Y. Xu, et al. “An experimental study on speech enhancement based on deep neural networks.” IEEE Signal Processing Letters, 2014, 21.1, pp. 65-68.
[18] X. Feng, Z. Yaodong and J. Glass. “Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition.” In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 1759-1763.
[19] T. Ishii, H. Komiyama, T. Shinozaki, Y. Horiuchi, and S. Kuroiwa, S. “Reverberant speech recognition based on denoising autoencoder”. In Proceedings of INTERSPEECH 2013, pp. 3512-3516.
bibitemsivasankaran2015 S. Sivasankaran, et al. “Robust ASR using neural network based speech enhancement and feature simulation.” IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) 2015, pp. 482-489.
[20] C. Weng, et al. “Recurrent deep neural networks for robust speech recognition.” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 5532-5536.
[21] M. Zhao, D. Wang, Z. Zhang, Z. and X. Zhang. “Music removal by denoising autoencoder in speech recognition”. IEEE Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 pp. 338-341.
[22] F. Seide, L. Gang and Y. Dong. “Conversational Speech Transcription Using Context-Dependent Deep Neural Networks.” Proceedings of INTERSPEECH 2011, pp. 437-440.
[23] M. Coto-Jiménez, J. Goddard-Close, J. and F.M. 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, pp. 354-361.
[24] X. Lu, et al. “Speech enhancement based on deep denoising autoencoder.” Proceedings of INTERSPEECH 2013, pp. 436-440.
[25] P. Scalart. “Speech enhancement based on a priori signal to noise estimation.” IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1996, pp. 629-632.
[26] F. Yuchen, et al. “TTS synthesis with bidirectional LSTM based recurrent neural networks.” Proceedings of INTERSPEECH 2014, pp. 1964-1968.
[27] H. Zen, and H. Sak. “Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis.” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, pp. 4470-4474.
[28] H. Sepp, and J. Schmidhuber. “Long short-term memory.” Neural computation 9.8 (1997), pp. 1735-1780.
[29] A. Graves, N. Jaitly, and A. Mohamed. “Hybrid speech recognition with deep bidirectional LSTM.” IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2013, pp. 273-278.
[30] A. Graves, S. Fernández, and J. Schmidhuber. “Bidirectional LSTM networks for improved phoneme classification and recognition.” International Conference on Artificial Neural Networks. Springer Berlin Heidelberg, 2005, pp. 799-804
[31] F.A. Gers, N. Schraudolph, and J. Schmidhuber, “Learning precise timing with LSTM recurrent networks.” Journal of machine learning research, 2002 3(Aug), pp. 115-143.
[32] V. Pascal, et al. “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.” The Journal of Machine Learning Research, 2010(11), pp. 3371-3408.
[33] S. Thomas, S.H.R. Mallidi, S. Ganapathy and H. Hermansky. “Adaptation transforms of auto-associative neural networks as features for speaker verification.” In Odyssey, 2012, pp. 98-104.
[34] K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink and J. Schmidhuber, J. “LSTM: A search space odyssey”. IEEE transactions on neural networks and learning systems, 28(10), 2017, pp. 2222-2232.
[35] J. Kominek and A.W. Black. “The CMU Arctic speech databases”. Fifth ISCA Workshop on Speech Synthesis, 2004, pp. 223-224.
[36] D. Erro, I. Sainz, E. Navas and I. Hernaez. “Improved HNM-based Vocoder for Statistical Synthesizers.” Proceedings of INTERSPEECH 2011, pp. 1809-1812.
[37] G. Beerends, et al. “Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part ii: psychoacoustic model.” Journal of the Audio Engineering Society 50.10, 2002, pp. 765-778.
[38] D. Klatt. “Prediction of perceived phonetic distance from critical-band spectra: A first step.” IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 1982, pp. 1278-1281.
[39] J.W. Tukey. “Comparing individual means in the analysis of variance.” Biometrics (1949): pp. 99-114.
[40] D.C. Montgomery. “Design and analysis of experiments.” John Wiley & sons, 2017.