Reconstructing fundamental frequency from noisy speech using initialized autoencoders

Marisol Zeledón-Córdoba, Joseline Sánchez-Solís, Marvin Coto-Jiménez

***Abstract*—In this paper, we present a new approach for fundamental frequency (f_0) detection in noisy speech, based on Long Short-term Memory Neural Networks (LSTM). f_0 is one of the most important parameters of human speech. Its detection is relevant in many speech signal processing areas and remains an important challenge for severely degraded signals. In previous references for f_0 detection in speech enhancement and noise reduction tasks, LSTM has been initialized with random weights, following a back-propagation through time algorithm to adjust them. Our proposal is an alternative for a more efficient initialization, based on the weights of an Auto-associative network. This initialization is a better starting point for the f_0 detection in noisy speech. We show the advantages of pre-training using objective measures for the parameter and the training process, with artificial and natural noise added at different signal-to-noise levels. Results show the performance of the LSTM increases in comparison to the random initialization, and represents a significant improvement in comparison with traditional initialization of neural networks for f_0 detection in noisy conditions.**

***Index Terms*—Deep Learning, Fundamental Frequency, LSTM, Neural Networks.**

I. Introducción

El procesamiento y análisis de las señales de habla degradadas con ruido ha sido considerado un tema de interés por diversos grupos de investigadores durante las últimas décadas. La razón principal de este interés se debe a la existencia de la distorsión que se presenta en entornos reales donde el habla es producida, registrada a través de micrófonos o transmitida y grabada por diversos medios. Esta distorsión puede afectar la calidad de los sistemas de comunicaciones o el desempeño de los sistemas de reconocimiento [1], [2].

Existen numerosos algoritmos desarrollados para mejorar las señales en estas condiciones adversas, los cuales se consideran exitosos si reducen el ruido de fondo que contamina la señal, si preservan o mejoran la calidad de la señal, o bien proporcionan una mejor detección de parámetros relevantes. La frecuencia fundamental (f_0) es un ejemplo de uno de estos parámetros, de importancia para identificar los mensajes y las características del hablante. Si bien existen algoritmos eficientes para la detección de este parámetro, ha sido señalada la necesidad de mejorar lo propuesto hasta ahora para la correcta estimación de f_0, especialmente en el caso del habla ruidosa [3].

Los métodos tradicionales para mejorar las señales y detectar más adecuadamente sus parámetros se basan en el procesamiento de señales, tales como el filtrado Wiener y la sustracción espectral, entre muchos otros, los cuales aprovechan parámetros estadísticos en distintos dominios.

Más recientemente, se ha presentado una nueva clase de algoritmos basados en deep learning, principalmente con redes neuronales profundas (DNN, por sus siglas en inglés). Estos se han presentado en diversas referencias [4], [5], reportando un éxito significativo y mejorando incluso el desempeño de los algoritmos basados en procesamiento de señales. El enfoque principal para las DNN es el mapeo de atributos espectrales de habla ruidosa en los atributos del habla limpia correspondiente.

Han surgido también nuevos tipos de redes neuronales con conexiones recurrentes (generalmente llamadas redes neuronales recurrentes o RNN, por sus siglas en inglés). En particular, la Red de Modelo de Memoria a Corto y Largo Plazo (LSTM, por sus siglas en inglés) ha tenido éxito sobre otros tipos de redes para reducir el ruido y las distorsiones reverberantes en las señales de habla. Los parámetros más comúnmente utilizados en estas aplicaciones para representar el espectro de las señales son los coeficientes de Mel-Frecuency Cepstrum (MFCC), que se derivan de la información de frecuencia.

En este trabajo, motivados por el éxito de las LSTM en la mejora del habla en diversas condiciones adversas, presentamos una forma de inicializar estas redes neuronales recurrentes para mejorar la detección de f_0 en el habla degradada con ruido.

Con las redes LSTM inicializadas de esta manera, mostramos los beneficios para la detección de f_0 en varios niveles de Relación Señal Ruido (SNR, por sus siglas en inglés), tanto en presencia de ruido artificial (Ruido Blanco), como natural (Ruido Babble).

A. Trabajo Relacionado

Para la detección de frecuencia fundamental en condiciones adversas, es decir, en presencia de distorsiones y de contaminación de distintos tipos, existe una diversidad de estudios basados en técnicas de procesamiento de señales. Éstas han explorado la información armónica de las señales en el dominio de la frecuencia o la periodicidad del dominio del tiempo [6], para estimar los valores de f_0. Recientemente, los algoritmos de *deep learning*, los cuales se basan principalmente en redes neuronales de múltiples capas, han sido usadas en numerosos trabajos relacionados a la reducción de ruido y detección de parámetros del habla [7], [8], [9].

El principal mecanismo para la mejora de las señales del habla usando algoritmos de aprendizaje profundo consiste en aplicar DNN como modelos de mapeo entre los parámetros del ruido en el habla hacia los parámetros del habla limpia

M. Zeledón-Córdoba, J. Sánchez-Solís y M. Coto-Jiménez laboran en el PRIS-Lab, Escuela de Ingeniería Eléctrica, Universidad de Costa Rica, San José, Costa Rica. correo electrónico: {marisol.zeledon, joseline.sanchezsolis, marvin.coto}@ucr.ac.cr

correspondientes [1], [10]. De esta manera, se obtienen versiones mejoradas de las señales donde es posible determinar con mayor precisión los parámetros deseados.

Algunas de estas aplicaciones de deep learning han superado a otros algoritmos de eliminación de ruido en la tarea de mejorar señales de habla que contienen ruidos diversos (naturales, artificiales) con varios niveles de SNR [11], [12].

Por ejemplo, en [13], las características del espectro del habla sintetizada son mejoradas empleando modelos como Redes de Creencia Profunda (DBNs, por sus siglas en inglés), y también Máquinas Restringidas de Boltzmann (RBM por sus siglas en inglés). Recientemente, el uso de RNN fue presentado, con la ventaja de que su estructura inherente parece lidiar mejor con la naturaleza dependiente del tiempo de los parámetros de las señales de habla. En estas referencias, el enfoque más común es mejorar los componentes espectrales del habla al asociarlos a sus equivalentes en la señal de referencia, sin ruido, reverberación u otra condición.

En la detección de f_0, las redes LSTM recientemente han superado otros múltiples algoritmos [14], [15]. En estos estudios ha quedado manifiesta la capacidad de los algoritmos de *deep learning* para mejorar las señales de habla degradadas, y de hacerlo de forma competitiva con respecto a otros algoritmos.

En este trabajo, se propone utilizar estos recursos de *deep learning* probados previamente en las referencias, y proporcionar una mejor etapa de inicialización para las redes neuronales, a manera de un pre-entrenamiento de las mismas. El proceso de entrenamiento de la red neuronal artificial, está tradicionalmente basado en una inicialización aleatoria de las conexiones internas, o pesos, que representan las uniones entre las entradas, las capas ocultas y las salidas. Después de la inicialización, se realiza el proceso de entrenamiento usual, empleando algoritmos tales como retro-propagación o retro-propagación a través del tiempo, dependiendo del tipo de red neuronal artificial.

Esta idea de inicialización de los pesos ha sido presentada en otros casos, inherente a modelos como las RBM [16], donde el proceso de inicialización es del tipo no supervisado. Los beneficios de estas etapas de pre-entrenamiento también han sido verificados en otras áreas de investigación, tales como la clasificación de música [17] y el reconocimiento facial [18]. Estos entrenamientos no supervisados presentan datos en la entrada de la red neuronal y actualizan los valores de los pesos sin comparar la salida con los datos correspondientes. Las técnicas semisupervisadas también han sido empleadas en aplicaciones similares [19], combinando la última etapa de datos sin etiquetar al inicializar los pesos de las redes neuronales.

En nuestra aproximación, el pre-entrenamiento es realizado con pares de datos presentados a la entrada y comparados con los valores esperados a la salida, de manera que se genera una red autoasociativa entrenada para aproximar la función identidad en el mapeo entre las entradas y las salidas. Una vez que esta función ha sido entrenada, los pesos correspondientes se convierten en los pesos iniciales de la red que se entrenará con pares de habla ruidosa y limpia para la posterior detección de la frecuencia fundamental.

B. Descripción del Problema

En aplicaciones relacionadas con reducción de ruido en señales de habla, usualmente se realizan procesamientos con la finalidad de mejorar la calidad de la señal, para convertirla en una versión cercana al habla sin este tipo de distorsiones. El proceso se puede modelar de la siguiente manera: Una señal ruidosa y se considera la suma de una señal limpia x, con un ruido d, es decir

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

En el dominio de la frecuencia, un equivalente se puede obtener al aplicar la Transformada de Fourier de tiempo corto, con la cual la señal se modela como:

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

donde k es el índice de la frecuencia y n es el índice del segmento de análisis. En la mayoría de métodos tradicionales de mejora de señales degradadas con ruido, $x(t)$ se considera no correlacionado con $d(t)$, de manera que se realiza una estimación de $X_k(n)$ a partir de los espectros de $y(t)$ y del estimado de $d(t)$.

Existen numerosos métodos para obtener $x(t)$ a partir de $y(t)$. Los que se basan en redes neuronales profundas pretenden una estimación de $x(t)$ directamente de un conjunto de datos que correspondan a pares de señal con ruido y señal sin ruido. Esta aproximación se realiza en la forma de una función $f(\cdot)$ que se ajusta durante el proceso de entrenamiento de las redes profundas. Esto se puede expresar como

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

donde $\hat{x}(t)$ se refiere a una aproximación de $x(t)$.

La reconstrucción de $x(t)$ a partir de los datos (es decir, la la precisión de la función de aproximación $f(\cdot)$), depende de factores como la cantidad de datos disponibles, del tipo de red neuronal empleada, su arquitectura e hiperparámetros.

Cuando se presentan los pares de datos correspondientes a señal con ruido y señal sin ruido en el entrenamiento, el conjunto de valores de las conexiones (o pesos) internos de la red se ajustan para obtener $f(\cdot)$. Los valores iniciales de los pesos se establecen tradicionalmente como números aleatorios, siguiendo alguna distribución de probabilidad. Cuando se presentan los datos, los valores se van actualizando hasta que se alcanza el criterio de paro, el cual se establece tradicionalmente como un máximo número de épocas, o bien un número preestablecido de épocas sin obtener mejora en el conjunto de validación.

En lugar de la inicialización con valores aleatorios de los pesos de la red, en este trabajo proponemos utilizar un pre-entrenamiento de la red en la forma de una función identidad (o memoria auto-asociativa), de manera que en el proceso de estimación de $f(\cdot)$ la red se encuentre en un estado inicial θ_A que sea más cercano y afín al mejor estado posible para la estimación de f_0 en las señales ruidosas.

En cuanto a este parámetro, la clasificación más general consta de dos categorías: sonoro y no sonoro. Los segmentos de habla sonoros son aquellos en los cuales se identifica un valor de frecuencia fundamental positivo (como en las vocales

y algunas consonantes líquidas), mientras que los segmentos no sonoros corresponden a valores cero de la frecuencia fundamental, más cercanos a breves segmentos de ruido. En esta segunda categoría se encuentran sonidos consonantes tales como /b/, /p/ y /s/.

Nuestra hipótesis principal es que θ_A es una mejor inicialización para las redes utilizadas que la realizada con valores aleatorios, principalmente en la detección de las fronteras sonoro/no sonoro presentes en el habla degradada con ruido.

El resto de este artículo está organizado de la siguiente manera: En la Sección II se presentan detalles de las redes LSTM, utilizadas para la detección de f_0 en las señales ruidosas. En la Sección III se describe la propuesta de implementación y los experimentos planteados para validarla. En la Sección IV se presentan y discuten los resultados, mientras que en la Sección V se presentan las conclusiones.

II. Redes Neuronales de Memoria a Corto y Largo Plazo

En la mejora del habla con ruido y la detección de f_0 en condiciones ruidosas, varios grupos de investigadores han experimentado con algoritmos de aprendizaje profundo. Las RNN [20], que incluyen retroalimentación hacia sí mismas y a otras neuronas en la misma capa, han logrado resultados particularmente buenos, en especial en el modelado de la dependencia natural de los parámetros del habla. Las redes LSTM se han presentado en [21] como una RNN extendida, con la capacidad de aprender a largo plazo relaciones entre los datos y almacenar información para intervalos largos o cortos de tiempo.

Entre las muchas implementaciones exitosas de LSTM se encuentran los sistemas automáticos de reconocimiento de voz, la síntesis de voz y la generación de letra manuscrita, donde los valores pasados de los parámetros son importantes para clasificar o realizar el mapeo requerido [22], [23].

Las redes LSTM tienen una estructura similar a las RNN básicas: un conjunto de unidades ingresa las secuencias $\mathbf{y} = (y_1, y_2, \ldots, y_T)$, y las secuencias vectoriales en las capas ocultas $\mathbf{h} = (h_1, h_2, \ldots, h_T)$ se calculan a través de un conjunto de pesos que están entre las entradas y las unidades ocultas, o entre unidades ocultas de las siguientes capas. Una descripción matemática detallada de las redes LSTM se puede encontrar en [21], [24]. En este trabajo seguimos la implementación descrita en [12].

En los enfoques de DNN para la detección de f_0, una red neuronal se entrena con varias entradas, desde donde f_0 se puede inferir. Se ha encontrado que los parámetros de la red utilizan datos de entrenamiento para minimizar el error de reconstrucción promedio, es decir, tener la salida $f(y)$ lo más parecida a la señal no degradada x [25], particularmente en el parámetro f_0.

Una de las arquitecturas recientes de las redes neuronales aplicadas para mejorar el habla ruidosa es el *autoencoder* de eliminación de ruido, que consta de dos partes: la primera es el codificador, donde un mapeo f transforma un vector de entrada y en una representación h en las capas ocultas. La segunda parte es el decodificador, donde se realiza un mapeo de la representación oculta en un vector $\hat{x}$.

Para la detección de f_0 en habla ruidosa, durante la etapa de entrenamiento, se presentan parámetros que fueron degradados con ruido en las entradas de los *autoencoders*, mientras que los atributos limpios correspondientes de la misma dimensión se convierten en salidas. El algoritmo de entrenamiento ajusta los parámetros de la red para aprender las complejas relaciones entre ellos y genera valores de f_0 que corresponde a los parámetros detectados del habla ruidosa.

III. Sistema Propuesto

Para detectar f_0 en las señales ruidosas de habla, el mapeo de f_0 se puede aprender directamente de los datos [26], con entrenamiento, validación y conjuntos de pruebas y procedimientos tradicionalmente definidos para algoritmos de aprendizaje automático. Para este propósito, en este trabajo se utilizan frases que consisten en habla a la que se agrega ruido, y la versión limpia correspondiente para entrenar las redes LSTM.

Los pesos de las redes LSTM son inicializados de tres formas:

- Aleatoriamente: Todas los pesos tienen números aleatorios al inicio de la primer época de entrenamiento. Esta es la forma más común en esta aplicación y las tareas de eliminación de ruido relacionadas. Consideramos este procedimiento como el sistema base.
- Autoasociativo: En ésta, las redes LSTM son entrenadas para presentar los mismos datos limpios en la entrada y salida en cada ventana. De esta forma, las redes aprenden la función identidad entre sus entradas y salidas. Después del entrenamiento, los pesos de las redes autoasociativas se convierten en los pesos iniciales de las redes LSTM correspondientes para la detección de f_0.
- Autoasociativo-ruidoso: Equivalente al anterior, pero utilizando parámetros de habla ruidosa para la inicialización de los pesos, en lugar de los parámetros del habla limpia utilizados en el punto anterior.

Estos tres sistemas se esquematizan en la Fig. 1. Los procesos de entrenamiento y prueba se realizaron para un nivel específico de SNR.

En todos los casos, se utilizó una arquitectura de tres capas ocultas, con 150, 100 y 150 neuronas en cada capa, de acuerdo con lo utilizado previamente en otras referencias [23], definidas después de un proceso de prueba y error. Al finalizar el entrenamiento de las redes autoasociativas, los pesos de las mismas fueron copiadas a las redes por entrenar, gracias a la coincidencia en arquitectura y número de conexiones.

A. Descripción de los Datos

En nuestros experimentos, usamos la base de datos producida en Carnegie Mellon University (CMU), descrita en [27], en particular la voz CMU-SLT. El conjunto de datos está fonéticamente balanceado, y fue originalmente diseñado para la investigación en síntesis de voz con la técnica de selección de unidades. Consiste en alrededor de 1150 frases seleccionadas de textos libres de derecho de autor del Proyecto Gutenberg.

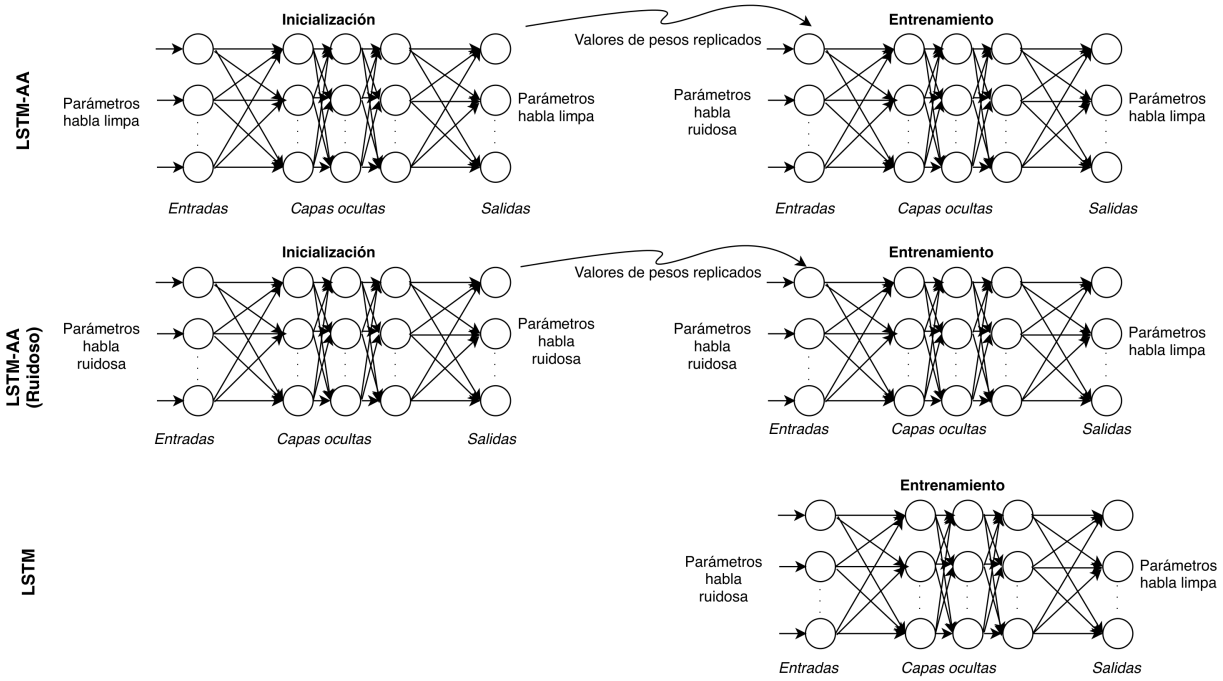

Fig. 1. Esquema de los tres sistemas comparados.

B. Extracción de Características

Los archivos de audio de la base de datos fueron re-muestreados a 16kHz, con la finalidad de extraer los parámetros usando el sistema Ahocoder. En este sistema, la frecuencia fundamental f_0^k (de valor cero en segmentos no sonoros), 39 MFCC, más un coeficiente de energía, son extraídos de cada ventana de 10 ms. Por lo tanto, cada ventana es representada por un vector de 41 dimensiones $V_k = \left[f_0^k, e^k, mfcc_k^1, \ldots, mfcc_k^{39}\right]$. Detalles sobre la extracción de parámetros y la reconstrucción de la forma de onda del sistema Ahocoder se pueden encontrar en [28].

C. Inicialización Autoasociativa

El procedimiento de inicialización se realizó usando 800 frases de la voz SLT. Para determinar la independencia de la inicialización del tipo de datos utilizados, se realizó tanto inicializando con parámetros de habla limpia, presentando los mismos parámetros a la entrada y salida de la red, como con habla ruidosa del nivel correspondiente en un segundo proceso. Se aplicaron los algoritmos habituales de retro-propagación a través del tiempo, y el criterio de parada fue 40 épocas desde el último mejor resultado, o a un máximo de 1000 épocas.

IV. Diseño Experimental

En esta sección se describe con mayor detalle el procedimiento experimental planteado para validar la propuesta de pre-entrenamiento de las redes neuronales. Este proceso se puede resumir en los siguientes pasos:

1) Generación de la base de datos con ruido: Ruido Blanco y Ruido Babble de distinta intensidad (de acuerdo con los cinco niveles SNR pre-establecidos) se agregó a cada archivo de audio de la base de datos. Los cinco niveles de SNR se definieron de manera que se contempla desde una afectación ligera de las señales, hasta degradación severa.
2) Extracción de características y correspondencias entradas/salidas: De cada uno de los archivos de audio, tanto limpios como degradados con ruido, se extrajo un conjunto de vectores con la parametrización de cada ventana, de 10 ms de duración. Para la tarea de detección de f_0, los parámetros que corresponden al habla ruidosa se presentan en la entrada de la red durante el entrenamiento, mientras que los correspondientes al habla limpia se presentan a la salida. Las inicializaciones correspondientes a los *autoencoders* se realizan presentando a éstos pares de habla limpia tanto a la entrada como a la salida, o bien pares de habla ruidosa, según sea el tipo de inicialización aplicada.
3) Entrenamiento y validación: Durante el proceso de entrenamiento, los pesos de las redes neuronales artificiales se ajustan, conforme se van presentando los pares (habla ruidosa, habla limpia) a la entrada y la salida. Para este proceso se utilizó un conjunto de 150 frases como validación.
4) Prueba: Para el conjunto de prueba, se seleccionaron aleatoriamente 50 frases. Estas frases no fueron parte del conjunto de entrenamiento ni el de validación, como una forma de garantizar independencia de los resultados.

La función de costo utilizada fue SSE (Suma de errores cuadráticos), y el optimizador fue descenso por gradiente, con tasa de aprendizaje $1e-5$. Para determinar la mejora en la eficiencia del procedimiento de pre-entrenamiento, se utilizaron las siguientes medidas objetivas, de uso frecuente

para evaluar detección de frecuencia fundamental [14]:

- DR (Razón de detección): Se evalúa sobre segmentos donde $f_0 > 0$, donde cada ventana de frecuencia fundamental se considera de estimación correcta si se encuentra bajo un 5% del valor correcto del parámetro. Es decir.

$$DR = \frac{N_{0.05}}{N_p} \times 100\%, \tag{4}$$

  donde $N_{0.05}$ representa el número de ventanas en las cuales el valor detectado de f_0 varía más del 5% del valor real, y N_p es el número total de ventanas.
- VDE (Error de decisión sonora): Indica el porcentaje de ventanas clasificadas incorrectamente en términos de sonora/no sonora. Se calcula de acuerdo con la ecuación:

$$VDE = \frac{N_{V\to U} + N_{U\to V}}{N} \times 100\%, \tag{5}$$

  donde $N_{V\to U}$ y $N_{V\to U}$ representan los errores de clasificación sonora a no sonora y viceversa.
- Suma de errores cuadráticos (SSE): Esta es una medida de uso frecuente para evaluar el proceso de entrenamiento de redes neuronales artificiales y el ajuste logrado con un conjunto particular. Se define como

$$\text{SSE}(\theta) = \sum_{n=1}^{T} (\mathbf{c_x} - \hat{\mathbf{c}_x})^2 \tag{6}$$

$$= \sum_{n=1}^{T} (\mathbf{c_x} - f(\mathbf{c_x}))^2, \tag{7}$$

  donde c_x es el valor conocido de las salidas y $\hat{c_x}$ es la aproximación realizada por la red. Esta medida se utiliza sobre todo con fines de ilustrar la mejora en el proceso de entrenamiento que provee la inicialización autoasociativa.

Finalmente, se utilizó el test estadístico de Friedman para determinar en cuáles conjuntos de resultados existe una diferencia estadísticamente significativa. Esto para permitir valorar no solamente una mejora en términos de su media, sino que ésta sea significativa al evaluar el conjunto de resultados en todo el conjunto de prueba.

Cabe destacar la posibilidad de aplicar métricas más elaboradas a un problema de esta naturaleza, tal como *Trend Similarity Evaluation* propuesta en [29], con la cual se puede plantear un análisis comparativo y extensión del presente.

V. Resultados y Discusión

Presentamos los resultados de la propuesta de este trabajo, relacionada con la inicialización de las redes y una comparación con las redes inicializadas aleatoriamente, utilizando en todo momento el algoritmo base de detección de f_0 del sistema Ahocoder. Los valores de f_0 que se reportan corresponden a $\log(f_0)$. Los cuatro sistemas que se considerarán son:

1) Ruidoso: detección de f_0 directamente desde el habla con ruido, provista con el algoritmo implementado en el sistema Ahocoder, basado en análisis armónico.
2) LSTM: detección de f_0 con la red LSTM inicializada con pesos aleatorios.
3) LSTM-AA: detección de f_0 con los pesos de la red LSTM inicializados desde la red auto-asociativa.
4) LSTM-AA (Ruidoso): detección de f_0 con los pesos de la red LSTM inicializados desde la red auto-asociativa, cuando ésta es entrenada inicialmente con parámetros de habla ruidosa.

La inicialización aleatoria se repitió tres veces, para medir y comparar las diferencias que se pueden presentar producto del proceso que inicia aleatoriamente. Cada proceso de entrenamiento de las redes neuronales fue acelerado utilizando GPU Nvidia, con una duración aproximada de seis horas. Los resultados de las redes que utilizan el pre-entrenamiento con red Auto-Asociativa se realizan solamente una vez, ya que no existe aleatoriedad en la inicialización ni en la presentación de los datos, por lo que los procesos de entrenamiento deben dar el mismo resultado en todos los casos, a diferencia de la inicialización aleatoria.

La Tabla I muestra los resultados del VDE para los tres sistemas y los cinco niveles de ruido blanco.

Con la excepción de SNR-10, la inicialización auto-asociativa propuesta presenta mejores valores de VDE en todos los niveles de SNR. La excepción puede explicarse en términos de los parámetros particularmente diferentes de la voz limpia utilizada en la inicialización de la red, que son muy diferentes de los de la señal ruidosa en las entradas. El resto de los resultados confirmaron que la inicialización auto-asociativa permite que las LSTM proporcionen mejores decisiones en cuanto a los segmentos sonoros/no sonoros del habla.

Los resultados de la medida DR se muestran en la Tabla II. Similar a la medida VDE, la inicialización auto-asociativa de la red LSTM presenta mejores resultados que la inicialización aleatoria, con la excepción de SNR-10. Las disminuciones significativas del valor DR en SNR0 y SNR-5 son consistentes con los mejores valores obtenidos de la medida VDE en estos niveles. Este resultado muestra cómo la inicialización propuesta beneficia a la detección de f_0 también en términos de la precisión de la detección de f_0.

El mejor rendimiento en VDE podría beneficiar los sistemas automáticos de reconocimiento de voz y la calidad perceptiva de las señales. Una ventaja adicional del LSTM-AA es el tiempo de entrenamiento, el cual hace más eficiente a la red y además cuenta con el menor error de SSE logrado, como el mostrado en la Fig. 2 con la evolución de la SSE en el conjunto de validación durante un procese de entrenamiento.

Si bien los resultados de la Fig.a 2 que corresponden al pre-entrenamiento con red Auto-Asociativa ya han tenido un proceso de entrenamiento previo, este proceso se realiza solo una vez y la red resultante se constituye en la inicialización de todas las demás utilizadas en el presente trabajo. Por esta razón, se estiman los beneficios en cuanto a menor número de épocas en esta figura, ya que es posible considerar un conjunto de redes pre-entrenadas como insumo para un nuevo conjunto de experimentos.

Finalmente, en la Tabla III se muestran los resultados de la prueba estadística del Test de Friedman, para determinar en qué casos existen diferencias estadísticamente significativas entre los resultados de la inicialización. Esta prueba fue realizada con un nivel de significancia de 0.05. Por ejemplo, para

TABLE I
COMPARACIÓN DE LOS RESULTADOS PARA LA MÉTRICA VDE EN EL CONJUNTO DE PRUEBA (50 FRASES). LSTM (An) REPRESENTAN LAS PRUEBAS CON INICIALIZACIÓN ALEATORIA. LSTM-AA SON LOS VALORES CON LA INICIALIZACIÓN PROPUESTA. LOS VALORES MENORES PORCENTUALES REPRESENTAN UN MEJOR RESULTADO.

Ruido Blanco						
SNR	Ruidoso	LSTM (A1)	LSTM (A2)	LSTM (A3)	LSTM-AA	LSTM-AA (Ruidoso)
-10	67.54%	5.35%	5.33%	5.44%	5.26% *	5.28%
-5	58.79%	3.81%	3.80%	3.78%	3.75%	3.74% *
0	25.20%	3.06%	3.05%	3.03%	3.02% *	3.02%*
5	9.24%	5.33%	5.33%	5.47%	5.30% *	5.42%
10	4.87%	7.18% *	7.58%	7.28%	7.18% *	7.19%
Ruido Babble						
SNR	Ruidoso	LSTM (A1)	LSTM (A2)	LSTM (A3)	LSTM-AA	LSTM-AA (Ruidoso)
-10	57%	10.39%	10.29%*	10.39%	10.43%	5.62%
-5	45.96%	7.91%	7.91%	7.77%	7.69% *	5.61%
0	27.25%	5.42%	5.37%	5.41%	5.36%*	5.62%
5	15.88%	5.33%	5.33%	5.47%	5.30%*	5.65%
10	9.88%	2.91%	3.03%	3.02%	3.00%*	5.63%

TABLE II
COMPARACIÓN DE LOS RESULTADOS PARA LA MÉTRICA DR EN EL CONJUNTO DE PRUEBA (50 FRASES). LSTM (An) REPRESENTAN LAS PRUEBAS CON INICIALIZACIÓN ALEATORIO. LSTM-AA SON LOS VALORES CON LA INICIALIZACIÓN PROPUESTA. LOS VALORES MENORES PORCENTUALES REPRESENTAN UN MEJOR RESULTADO.

Ruido Blanco						
SNR	Ninguno	LSTM (A1)	LSTM (A2)	LSTM (A3)	LSTM-AA	LSTM-AA (Ruidoso)
-10	100%	8.13% *	14.92%	14.01%	8.86%	16.59%
-5	82.63%	8.58%	8.41%	8.53%	4.64% *	7.58%
0	47.40%	8.59%	7.58%	5.92%	4.26% *	10.50%
5	14.11%	2.92%	3.79%	10.35%	2.26% *	9.61%
10	7.19%	1.02% *	1.16%	3.77%	1.16%	4.48%
Ruido Babble						
SNR	Ninguno	LSTM (A1)	LSTM (A2)	LSTM (A3)	LSTM-AA	LSTM-AA (Ruidoso)
-10	100%	15.48%	15.34%	15.33%	9.27%*	11.05%
-5	85.93%	15.72%	14.79%	16.23%	9.68%*	9.72%
0	37.45%	13.69%	6.92%	8.10%	5.89%*	11.95%
5	22.12%	2.92%	3.79%	10.35%	2.36%*	10.52%
10	11.73%	8.81%	8.56%	5.22%	4.21% *	10.84%

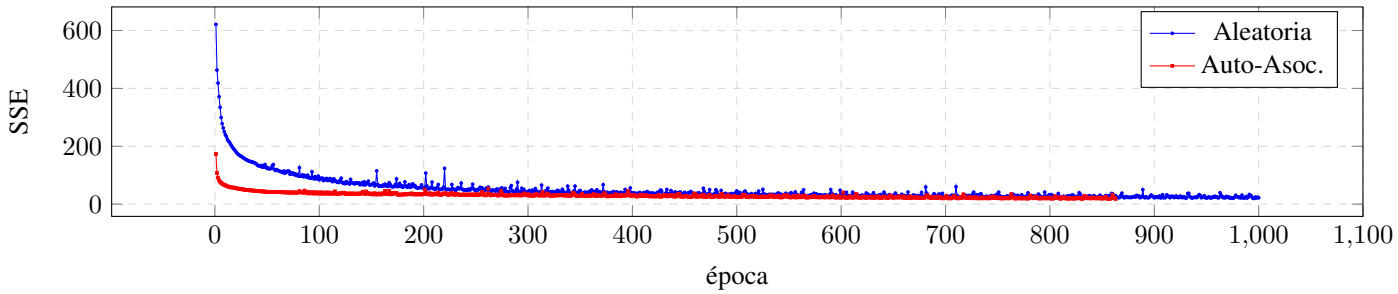

Fig. 2. Evolución del error SSE en cada época para la red LSTM con inicialización aleatoria (el mejor de los tres procesos realizados) y el proceso de pre-entrenamiento con red auto-asociativa, caso SNR 0. El valor SSE más bajo significa mejores resultados, mientras que la menor cantidad de épocas necesarias para alcanzar el SSE mínimo representa un entrenamiento más eficiente.

el caso de la métrica VDE, se puede observar cómo los beneficios de la inicialización autoasociativa representan mejoras estadísticamente significativas con respecto a la inicialización aleatoria. Si bien esta mejora no se da en todos los casos, los resultados permiten afirmar la conveniencia de usar un proceso de inicialización, dado que el procedimiento aleatorio puede llegar a resultados significativamente por debajo de los proporcionados con la inicialización autosociativa.

Las diferencias significativas son aún más notorias en el caso de la métrica DR. Aquí, como se puede observar en la Tabla III, tanto para el caso de ruido Blanco como ruido Babble, la inicialización aleatoria con mucha frecuencia presenta resultados significativamente distintos a nuestra propuesta. Esta información valida la necesidad de aplicar el proceso

TABLE III
SIGNIFICANCIA ESTADÍSTICA EN LA MEJORA DEL VALOR DE DR Y VDE. LSTM (An) REPRESENTAN LAS PRUEBAS CON INICIALIZACIÓN ALEATORIA. LSTM-AA SON LO VALORES CON LA INICIALIZACIÓN PROPUESTA. CON * SE INDICA EL MEJOR RESULTADO, Y CON LA PALABRA SÍ/NO SE INDICAN AQUELLOS QUE NO DIFIEREN SIGNIFICATIVAMENTE. ENTRE PARÉNTESIS SE REPORTA EL P-VALOR DEL TEST DE FRIEDMAN.

			VDE: R. Blanco			
SNR	None	LSTM (A1)	LSTM (A2)	LSTM (A3)	LSTM-AA	LSTM-AA (ruidoso)
-10	Sí (p=0.00)	No (p=0.53)	Sí (p=0.00)	Sí (p=0.02)	Mejor caso	Sí (p=0.00)
-5	Sí (p=0.00)	No (p=0.13)	No (p=0.21)	Sí (p=0.00)	Sí (p=0.00)	Mejor caso
0	Sí (p=0.00)	No (p=0.73)	No (p=0.87)	No (p=0.52)	Mejor caso	No (p=0.64)
5	Sí (p=0.00)	No (p=0.28)	No (p=0.42)	Sí (p=0.00)	Mejor caso	No (p=1.00)
10	Sí (p=0.00)	No (p=0.86)	No (p=0.58)	No (p=0.17)	Mejor caso	No (p=0.87)
			VDE: R. Babble			
SNR	None	LSTM (A1)	LSTM (A2)	LSTM (A3)	LSTM-AA	LSTM-AA (ruidoso)
-10	Sí (p=0.00)	No (p=0.33)	No (p=0.39)	No (p=0.77)	Mejor caso	Sí (p=0.00)
-5	Sí (p=0.00)	No (p=0.13)	Sí (p=0.01)	No (p=0.37)	Mejor caso	Sí (p=0.00)
0	Sí (p=0.00)	No (p=0.46)	No (p=0.74)	No (p=0.54)	Mejor caso	No (p=0.41)
5	Sí (p=0.00)	No (p=0.87)	No (p=0.45)	No (p=0.28)	Mejor caso	Sí (p=0.00)
10	Sí (p=0.00)	No (p=0.16)	Sí (p=0.04)	Sí (p=0.01)	Mejor caso	Sí (p=0.00)
			VR: R. Blanco			
SNR	None	LSTM (A1)	LSTM (A2)	LSTM (A3)	LSTM-AA	LSTM-AA (ruidoso)
-10	Sí (p=0.00)	Mejor caso	Sí (p=0.00)	Sí (p=0.00)	Sí (p=0.00)	No (p=0.46)
-5	Sí (p=0.00)	Sí (p=0.00)	Sí (p=0.00)	Sí (p=0.00)	Mejor caso	No (p=0.77)
0	Sí (p=0.00)	Sí (p=0.00)	Sí (p=0.00)	Sí (p=0.00)	Mejor caso	Sí (p=0.00)
5	Sí (p=0.00)	Sí (p=0.00)	Sí (p=0.00)	Sí (p=0.00)	Mejor caso	Sí (p=0.00)
10	Sí (p=0.00)	Mejor caso	No (p=0.38)	Sí (p=0.00)	No (p=0.38)	Sí (p=0.00)
			VR: R. Babble			
SNR	None	LSTM (A1)	LSTM (A2)	LSTM (A3)	LSTM-AA	LSTM-AA (ruidoso)
-10	Sí (p=0.00)	Sí (p=0.00)	Sí (p=0.00)	Sí (p=0.00)	Mejor caso	No (p=0.28)
-5	Sí (p=0.00)	Sí (p=0.00)	Sí (p=0.00)	Sí (p=0.00)	Mejor caso	No (p=0.28)
0	Sí (p=0.00)	Sí (p=0.00)	Sí (p=0.00)	Sí (p=0.00)	Mejor caso	Sí (p=0.00)
5	Sí (p=0.00)	Sí (p=0.00)	Sí (p=0.00)	Sí (p=0.00)	Mejor caso	Sí (p=0.00)
10	Sí (p=0.00)	Sí (p=0.00)	Sí (p=0.00)	Sí (p=0.00)	Mejor caso	Sí (p=0.00)

de inicialización para aplicaciones semejantes de mapeo de parámetros basados en arquitecturas como *autoencoders*.

Para otras aplicaciones relacionadas con redes LSTM utilizadas en problemas de mejora de señales, u otros donde se requieran mapeos entre parámetros de naturaleza semejante, el proceso de pre-entrenamiento se puede generalizar con la utilización de los valores en el conjunto de entrenamiento, tanto a la entrada como a la salida. De esta manera, se tiene una aproximación de la función identidad en la red LSTM, en la que posteriormente se realiza el proceso de entrenamiento de forma usual.

VI. CONCLUSIONES

En este trabajo se presentó una propuesta para la inicialización de redes LSTM con el objetivo de mejorar la detección de f_0 en señales de habla ruidosa. Esta inicialización está basada en los pesos de una red autoasociativa. Se realizó una comparación en cinco niveles de ruido Blanco y de ruido Babble, utilizando dos métricas conocidas para la evaluación en la precisión de la detección de f_0.

Las redes LSTM presentan una alta capacidad para detectar f_0 en condiciones de altos niveles de ruido, donde la frecuencia fundamental es totalmente indetectable para los algoritmos aplicados, en particular SNR-10 donde se obtuvo hasta un 100% de error sin la aplicación de las redes neuronales. Nuestra propuesta para la inicialización autoasociativa funcionó mejor en términos de un tiempo más eficiente de entrenamiento, y menores valores de la tasa de detección de segmentos sonoros y no sonoros, y error de decisión de voz, en contraste con la inicialización aleatoria. A pesar de requerir el proceso de inicialización previa, los beneficios de la propuesta pueden constatarse cuando se realiza un número considerable de pruebas, para las cuales solamente se realiza una sola inicialización con la cual se reduce el tiempo de entrenamiento del resto de redes neuronales.

La premisa principal para aplicar la propuesta fue contar con una mejor aproximación para el mapeo realizado desde el habla ruidosa hacia los parámetros del habla limpia. Para validar los resultados se utilizó una prueba estadística, con la cual se puede afirmar que los beneficios de la propuesta tienen significancia estadística en gran cantidad de casos, lo que la hace un procedimiento más seguro en la aplicación probada.

AGRADECIMIENTOS

El presente trabajo ha sido desarrollado gracias al apoyo de la Universidad de Costa Rica, proyecto 322-B9-105.

REFERENCES

[1] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "Feature enhancement by deep lstm networks for asr in reverberant multisource environments," *Computer Speech & Language*, vol. 28, no. 4, pp. 888–902, 2014.

[2] D. Bagchi, M. I. Mandel, Z. Wang, Y. He, A. Plummer, and E. Fosler-Lussier, "Combining spectral feature mapping and multi-channel model-based source separation for noise-robust automatic speech recognition," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 496–503.

[3] K. Wu, D. Zhang, and G. Lu, "ipeeh: Improving pitch estimation by enhancing harmonics," *Expert Systems with Applications*, vol. 64, pp. 317–329, 2016.

[4] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, and C.-H. Lee, "Robust speech recognition with speech enhanced deep neural networks," in *15th Annual Conference of the International Speech Communication Association*, 2014.

[5] K. Han, Y. He, D. Bagchi, E. Fosler-Lussier, and D. Wang, "Deep neural network based spectral feature mapping for robust speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] K. Han and D. Wang, "Neural network based pitch tracking in very noisy speech," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 2158–2168, 2014.

[7] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7398–7402.

[8] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4277–4280.

[9] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7596–7599.

[10] M. Coto-Jiménez and J. Goddard-Close, "Lstm deep neural networks postfiltering for enhancing synthetic voices," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 32, no. 01, p. 1860008, 2018.

[11] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, 2012.

[12] M. Coto-Jiménez, J. Goddard-Close, and F. Martínez-Licona, "Improving automatic speech recognition containing additive noise using deep denoising autoencoders of lstm networks," in *International Conference on Speech and Computer*. Springer, 2016, pp. 354–361.

[13] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets." in *Interspeech*, 2013, pp. 369–372.

[14] B. Liu, J. Tao, D. Zhang, and Y. Zheng, "A novel pitch extraction based on jointly trained deep blstm recurrent neural networks with bottleneck features," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 336–340.

[15] I. Thoidis, L. Vrysis, K. Pastiadis, K. Markou, and G. Papanikolaou, "Investigation of an encoder-decoder lstm model on the enhancement of speech intelligibility in noise for hearing impaired listeners," in *Audio Engineering Society Convention 146*. Audio Engineering Society, 2019.

[16] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2011.

[17] A. Van Den Oord, S. Dieleman, and B. Schrauwen, "Transfer learning by supervised pre-training for audio-based music classification," in *Conference of the International Society for Music Information Retrieval (ISMIR 2014)*, 2014.

[18] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International Conference on Machine Learning*, 2014, pp. 647–655.

[19] K. Veselỳ, M. Hannemann, and L. Burget, "Semi-supervised training of deep neural networks," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 267–272.

[20] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "Tts synthesis with bidirectional lstm based recurrent neural networks," in *15th Annual Conference of the International Speech Communication Association*, 2014.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 273–278.

[23] M. Coto-Jiménez, "Improving post-filtering of artificial speech using pre-trained lstm neural networks," *Biomimetics*, vol. 4, no. 2, p. 39, 2019.

[24] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with lstm recurrent networks," *Journal of Machine Learning Research*, vol. 3, no. Aug, pp. 115–143, 2002.

[25] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[26] L.-H. Chen, T. Raitio, C. Valentini-Botinhao, Z.-H. Ling, and J. Yamagishi, "A deep generative architecture for postfiltering in statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 11, pp. 2003–2014, 2015.

[27] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[28] D. Erro, I. Sainz, I. Saratxaga, E. Navas, and I. Hernáez, "Mfcc+ f0 extraction and waveform reconstruction using hnm: preliminary results in an hmm-based synthesizer," *Proc. FALA*, pp. 29–32, 2010.

[29] J. P. Arias, N. B. Yoma, and H. Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech communication*, vol. 52, no. 3, pp. 254–267, 2010.

Marisol Zeledón-Córdoba is currently a researcher at PRIS-Lab, in the School of Electrical Engineering at the University of Costa Rica. She has been involved in speech technologies and speech enhancement projects based on deep learning methods. Her main research interests are pattern recognition and speech technologies. marisol.zeledon@ucr.ac.cr

Joseline Sánchez-Solís is currently a researcher at the PRIS-Lab, in the School of Electrical Engineering at the University of Costa Rica, where she works on projects related to speech technologies and sound perception. Her main research interests are intelligent systems and telecommunications, as well as the applications of speech technologies. joseline.sanchezsolis@ucr.ac.cr

Marvin Coto-Jiménez received a B.S. degree in Electrical Engineering and M.Sc. in Mathematics at the University of Costa Rica, in Costa Rica. Also, he received an M.Sc. and Ph.D. in Information Technologies at the Metropolitan Autonomous University, in Mexico City. His main research interests are pattern recognition and intelligent systems applied to speech technologies. marvin.coto@ucr.ac.cr