Ingeniare. Revista Chilena de Ingeniería
ISSN: 0718-3291
facing@uta.cl
Universidad de Tarapacá
Chile

San Juan, Enrique; Jamett, Marcela; Kaschel, Héctor; Sánchez, Luis
Sistema de reconocimiento de voz mediante wavelets, predicción lineal y redes backpropagation
Ingeniare. Revista Chilena de Ingeniería, vol. 24, núm. 1, enero, 2016, pp. 8-17
Universidad de Tarapacá
Arica, Chile

Sistema de reconocimiento de voz mediante wavelets, predicción lineal y redes backpropagation

Wavelets, linear prediction and backpropagation networks for speech recognition system

Enrique San Juan[1] Marcela Jamett[1] Héctor Kaschel[1] Luis Sánchez[1]

Recibido 19 de mayo de 2014, aceptado 12 de mayo de 2015
Received: May 19, 2014 Accepted: May 12, 2015

RESUMEN

En el presente trabajo se muestra un sistema que combina la utilización de Transformadas de Wavalet (WT, *Wavalet Transform*), Coeficientes de Predicción Lineal (LPC, *Linear Predictive Coding*) y Redes Neuronales Artificiales (ANN, *Artificial Neural Networks*), para reconocer vocales y sílabas en forma independiente del hablante. Usando esta estructura, se propone un software automatizado que, a través de una interfaz, permite a usuarios con dificultades de audición o ausencia total de ésta, la posibilidad de emplearlo en primera instancia como una herramienta inicial de apoyo para el aprendizaje de sílabas. En una primera etapa se incorpora un número reducido de sílabas, en especial las que podrían tener más dificultad para su identificación. Posteriormente se puede ir incorporando una mayor cantidad de sílabas, de tal forma que el sistema pueda ir creciendo. La adición de nuevas sílabas, permitirá luego (a través de la segmentación de palabras en sílabas), la implementación de un sistema mayor para identificar palabras y por consiguiente el aprendizaje del lenguaje hablado.

Palabras clave: Coeficientes LPC, reconocimiento de voz, reconocimiento de patrones, transformada wavelet (WT), redes neuronales artificiales (ANN).

ABSTRACT

A system that combines the use of Wavalet Transforms (WT), Linear Prediction Coefficients (LPC) and Artificial Neural Networks (ANN) are shown in this paper. Vowels and syllables recognition speaker independent can be made. Using this structure, automated software is proposed, which through an interface, allows users with hearing difficulties or total absence of this, the possibility of using it primarily as an initial tool to support learning of syllables. In the first stage, a small number of syllables is incorporated, especially those that may have more difficulty in their identification. Subsequently, it can be incorporated a larger number of syllables, then it can be growing. Adding new syllables, would allow (through segmenting words into syllables) a greater deployment system, for identifying complete words and therefore the spoken language learning.

Keywords: LPC coefficients, speech recognition, pattern recognition, wavelet transform (WT), artificial neural networks (ANN).

[1] Departamento de Ingeniería Eléctrica. Universidad de Santiago de Chile. Av. Ecuador 3519, Estación Central. Santiago, Chile. E-mail: enrique.sanjuan@usach.cl; marcela.jamett@usach.cl; hector.kaschel@usach.cl; luis.sanchez@usach.cl

INTRODUCCIÓN

El trastorno de audición (TA) desde el punto de vista médico se define como: "La falta total o casi total de la audición, que se manifiesta desde el nacimiento o por lo menos antes de que un niño comience a hablar" [1]. Tomando en cuenta esta definición, el TA es una de las pérdidas sensoriales más perjudiciales para el desenvolvimiento social de un ser humano, ya que no solamente impide que una persona pueda escuchar, sino que también le puede imposibilitar hablar correctamente o no hablar [2]. Al quedarse sin audición, un niño no aprende el sistema del lenguaje hablado, luego no puede adquirir las herramientas necesarias para comunicarse adecuadamente, pudiéndose afectar en forma significativa su desarrollo intelectual y su inserción en la sociedad.

Las únicas herramientas que en la actualidad se utilizan para enseñar a personas con TA, son la lectura de labios y el lenguaje de señas. El desarrollo de herramientas multimedia para enseñar el lenguaje hablado, es por tanto una necesidad social de gran importancia. Se considera el estudio y desarrollo de una herramienta que al aplicarse experimentalmente, pueda ayudar a personas con TA a entrenarse en el aprendizaje lenguaje hablado.

Con el propósito de obtener muestras variadas, se tomó hablantes de lengua española de distintos géneros y edades, incluyendo niños. Todos ellos sin TA y procedentes de Chile (país en que reside el estudio).

En este artículo se muestran los resultados para las 5 vocales del habla hispana y además 4 fonemas característicos: "ya", "te", "pi", "bu". En ellos están presentes cualidades distintivas: consonantes cortas, largas, sonoras y no sonoras, ruidosas, entre otras.

ESTADO DEL ARTE

Desde hace varias décadas, los coeficientes LPC son una de las herramientas más utilizadas para el análisis de la señales y en particular de la voz [3-4] y que gracias al aumento de la capacidad de procesos de los sistemas computacionales en los últimos años, ha posibilitado que las redes neuronales y la transformada wavelet se incorporen también en su estudio.

Algunas aplicaciones interesantes se pueden ver en [5-6], en las que se presentan sistemas bimodales (audio y visual) para reconocimiento de voz, en particular para los idiomas inglés y malayo. Estos trabajos se centraron en la aplicación bajo ciertas restricciones (de bajo ruido) para hablantes comunes (sin TA); dando énfasis en el uso de técnicas de filtrado, extracción de características y clasificación.

Las herramientas matemáticas utilizadas para el procesamiento de señales de voz son tan diversas como de multiuso: en el trabajo de [7] se presenta el uso de wavelets como filtros previos al modelado de sistemas, pues se reconoce que el ruido ambiental es un problema complejo en los procesos de reconocimiento de voz. En este sentido, la combinación de wavelets y redes neuronales [8-9] se propone como alternativa para lograr una mejor robustez en el reconocimiento de voz frente a situaciones ruidosas (normales).

En [3], Paul, Das y Kamal proponen una combinación de LPC y redes neuronales con el fin de realizar el procesamiento de la señal con detección de inicio, ventaneo, filtrado, coeficientes LPC y ceptrales, para construir los códigos que posteriormente son tomados como patrones que una RN debe reconocer. Este trabajo se aplicó al idioma bengalí.

El uso de las técnicas de LPC, wavelets y redes neuronales para reconocimiento del habla se presentan en [10], específicamente para el idioma Inglés. Mientras que en [11] se plantea el uso de estas mismas metodologías para el idioma Árabe, en particular, para reconocimiento de vocales.

Las aplicaciones anteriormente citadas, se centran en el uso por parte de personas sin restricciones auditivas y/o del habla. Luego en [12] se presenta un estudio para hablantes con disartria (en el idioma español – mexicano), lo que constituye una ayuda para mejorar el habla de estas personas. Se usa el algoritmo de Viterbi para el reconocimiento del habla.

En el ámbito de las aplicaciones para personas con discapacidad, se encuentran los trabajos de [13-14] en los que se desarrollan robots de asistencia a discapacitados motores con reconocimiento de instrucciones de voz. Por otra parte, en [15] se plantea un mecanismo de reconocimiento de habla para personas con patologías en la misma, ya sea por razones neuronales y/o de la voz.

Finalmente, en [16] se propone un sistema de identificación gestual sobre la base de un sistema de video que permite reconocer letras. En esta aplicación se vislumbran futuras aplicaciones orientadas a posibilitar la interacción de personas sordomudas con la población, en general.

PROPUESTA

La discusión anteriormente presentada, da pié a la actual propuesta, que consiste en la combinación de técnicas LPC (Linear Predictive Coding), WT (Wavelet Transform) y ANN (Artificial Neural Networks) para el desarrollo de un sistema de apoyo al aprendizaje de pronunciación de vocales y sílabas del idioma español. Ésta constituye una herramienta virtual en el reconocimiento de voz, independiente del hablante. Lo que la diferencia de los estudios anteriormente mencionados es el tipo de público al cual va dirigido: las personas con TA.

Cabe aclarar que a pesar de que los coeficientes ceptrales en la escala de frecuencias de Mel son más robustos que los coeficientes LPC, porque adaptan las frecuencias de fonemas a la manera que el oído humano percibe los sonidos, para efectos de cálculo, la cantidad de coeficientes de Mel es superior la cantidad de LPC [17-18]. Por otra parte en [18] se observa que el coeficiente de correlación de Pearson es muy cercano entre ambos. De esta manera, para optimizar el uso del sistema integrado por LPC, WT y ANN, se elige usar LPC en lugar de coeficientes de Mel.

Se presenta el desarrollo de un software que modela el procesamiento digital de la voz mediante la combinación de tres herramientas: la WT, los coeficientes LPC y las redes neuronales BP (backpropagation).

Considerando las 5 vocales del habla española y un número reducido de sílabas, en especial las que podrían tener más dificultad para su identificación, se encuentran cualidades distintivas: consonantes cortas, largas, sonoras y no sonoras, ruidosas, entre otras (para este paper se presentan los resultados para 4 sílabas).

En etapa de desarrollo está la incorporación de una mayor cantidad de sílabas, de tal forma que el sistema pueda evolucionar a un sistema que reconozca palabras, mediante la segmentación de las mismas en sílabas.

El software permite a usuarios con dificultades de audición o ausencia total de ésta, emplearlo como una herramienta inicial de apoyo para el aprendizaje de vocales y sílabas. El usuario con TA, de manera autónoma o asistida, digita la vocal o sílaba que quiere aprender, mediante el entrenamiento de su pronunciación. Luego, la selecciona a través de una interfaz gráfica amigable, donde se le entrega instrucciones con lenguajes de señas y videos con el movimiento de los labios al pronunciar la misma. El usuario realiza la pronunciación luego de una invitación a grabar. El programa adquiere esta información y la procesa con el modelo planteado: WT, LPC y ANN, le indica el nivel de cercanía que tuvo en la pronunciación en forma porcentual y cualitativa usando la escala “muy bien”, “bien”, “regular” o “lejano”. En la Figura 1 se muestra un ejemplo de este programa en funcionamiento.

Figura 1. Interfaz gráfica del software de entrenamiento.

En cuanto a las técnicas de procesamiento, la red neuronal utilizada es la BPNN (Backpropagation Neural Network), seleccionada porque permite generalizar modelos de muestras que difieren entre sí; luego se emplean en la etapa de aprendizaje supervisado una vez que han sido preprocesadas por la técnica LPC, obteniendo sus respectivos coeficientes.

Por otra parte, la WT es una forma de análisis de una señal en tiempo y frecuencia. Como su nombre lo indica, descompone una señal en pequeñas ondas de menor resolución y distintas frecuencias. Es así como la WT descubre detalles importantes a lo largo de la duración de la señal y que son reflejadas en las

pequeñas ondas, pudiendo localizarlas en la señal original, a diferencia de la transformada de Fourier que detecta la frecuencia de la perturbación pero no su localización en el tiempo.

La técnica de WT es empleada en este trabajo para limpiar la señal del ruido, disminuir el número de muestras necesarias y el procesamiento posterior de la misma.

TÉCNICAS DIGITALES PARA EL ANÁLISIS DE LA VOZ UTILIZADAS

A continuación se realiza una descripción de los principales modelos empleados en este trabajo.

Predicción lineal en el dominio del tiempo
En predicción lineal es ampliamente utilizado el modelo para todo polo, conocido como modelo autorregresivo. En este modelo, la señal, s_n, se expresa mediante una combinación lineal entre sus valores pasados, s_{n-k}

$$\tilde{s}_n = -\sum_{k=1}^{p} a_k s_{n-k} \qquad (1)$$

Los pesos a_k son conocidos como parámetros LPC y se calculan a partir de la minimización del error total al cuadrado, dado por la suma de las diferencias entre el valor real y el predicho para cada secuencia de la sucesión $\tilde{s}_n$. Al aplicar este criterio se obtiene un conjunto de ecuaciones de p incógnitas. Para la resolución del sistema de ecuaciones planteado se utiliza el método de autocorrelación, mediante el algoritmo de Levison-Durbin [18]. Dichos parámetros son calculados para ventanas de voz, en donde su espectro se considera estacionario para tramos cortos de tiempo, considerados entre 10 ms y 40 ms. En donde los a_k son una forma de representación paramétrica de la señal y contienen la información relevante de la misma.

Red Neuronal Backpropagation
El funcionamiento de una red BP consiste básicamente en un aprendizaje de un conjunto predefinido de pares de entradas y salidas, empleando un ciclo propagación-adaptación de dos fases.

La ventaja del uso de la red BP radica en su capacidad de adaptar los pesos de sus elementos en capas intermedias, de manera de aprender la relación que existe entre un conjunto de patrones dados como ejemplo (entradas) y las salidas de la red. Luego, aplicada a nuevos vectores de entrada con ruido o incompletos, entrega una salida activa si la nueva entrada es parecida a las precedentes. Esta propiedad se conoce como generalización.

Dado que el algoritmo utilizado es el backpropagation, las funciones de activación deben ser continuas para que sean diferenciables. La función utilizada es del tipo sigmoidal por simplicidad [19].

En la práctica, existe una regla empírica que rige la cantidad de ejemplos a utilizar para el proceso de entrenamiento, para las pruebas y para la validación [19]. Así:

- El 70% debe ser para entrenamiento
- El 20% debe ser para pruebas
- El 10% debe ser para validación

No existe una ley matemática que permita calcular cuántos ejemplos deben ser utilizados para el entrenamiento de las redes neuronales, sólo se poseen reglas empíricas que estiman la cantidad mínima de ejemplos para asegurar un buen funcionamiento de una ANN. En [20] se propone la siguiente fórmula para lograr este propósito

$$\begin{aligned} N° \; de \; ej. &\geq 10 N_{tp} \\ N_{tp} &= (N_e + 1) N_{co} + (N_{co} + 1) N_s \end{aligned} \qquad (2)$$

donde N_{tp} es el número total de pesos w, N_e la cantidad de neuronas de la capa de entrada, N_{co} la cantidad de neuronas de la capa oculta y N_s la cantidad de neuronas de la capa de salida.

Transformada Wavelet
La WT es una forma de representar una señal compleja de manera simple [21], lo cual la convierte en una herramienta útil desde el punto de vista práctico. El análisis mediante wavelets consiste en dividir una señal en un determinado número de ondas o combinaciones lineales de señales de duración finita resultantes de la traslación y escalado de una función wavelet madre [22].

Se distinguen dos tipos de WT; la transformada wavelet continua (CWT, Continuous Wavelet Transform) y la transformada discreta (DWT,

Discrete Wavelet Transform). En la CWT una señal $f(t)$ de tiempo continuo, se representa mediante una expansión de términos o coeficientes que son proporcionales al producto interno entre la señal y las diferentes versiones escaladas y trasladadas de una función prototipo $\psi(t)$, mejor conocida como wavelet madre [23].

PLANTEAMIENTO DEL MODELO

En la Figura 2 se muestra el modelo desarrollado para el proceso de reconocimiento de vocales y sílabas. En primer lugar, el usuario con TA pronuncia una vocal o una sílaba, para luego capturar la señal de voz.

Posteriormente se procede a filtrar el ruido proveniente de perturbaciones indeseadas sobre la señal; por ejemplo de hardware, ruido externo o ambiental.

El objetivo de este procedimiento es llevar a cabo un procesamiento posterior más confiable y eficiente de la señal. Este filtrado se realiza a través del análisis de multiresolusión wavelet que permite descomponer la señal original en distintos niveles de resolución; los cuales son ponderados por los k_i, los que corresponden a los pesos de participación de cada detalle, obtenidos mediante un *toolbox* de WT. La WT madre utilizada es la Daubechies 6, donde la cantidad de niveles y el tipo de WT se determinan empíricamente, luego de haber probado con distintas familias y niveles.

Posterior a la reducción del ruido, se reconstruye la señal a partir de s_d, luego se identifica si se trata de una vocal o de una sílaba, lo que es seleccionado por el usuario a través de la interfaz gráfica. Si es el caso de una vocal, se aplica nuevamente la WT y se selecciona el detalle d_3, el cual entrega la mejor representación de la señal, esta selección es arbitraria de acuerdo a resultados empíricos, filtrada de ruido y que además ha disminuido su resolución (la cantidad de muestras).

A continuación, se realiza un análisis LPC, ingresando 7 coeficientes, de acuerdo al siguiente razonamiento: para un tracto vocal normal (17 cm), hay un promedio de un formante por kHz de ancho de banda. Un formante requiere 2 polos complejos conjugados, necesitándose dos coeficientes predictores [24], luego, de acuerdo a [25], los 4 primeros formantes representan las principales características de la voz, lo que conlleva al uso de 8 LPC. Por otra parte, estos se normalizan respecto al primero, por lo que es suficiente tomar a partir del segundo, así, 7 LPC son suficientes.

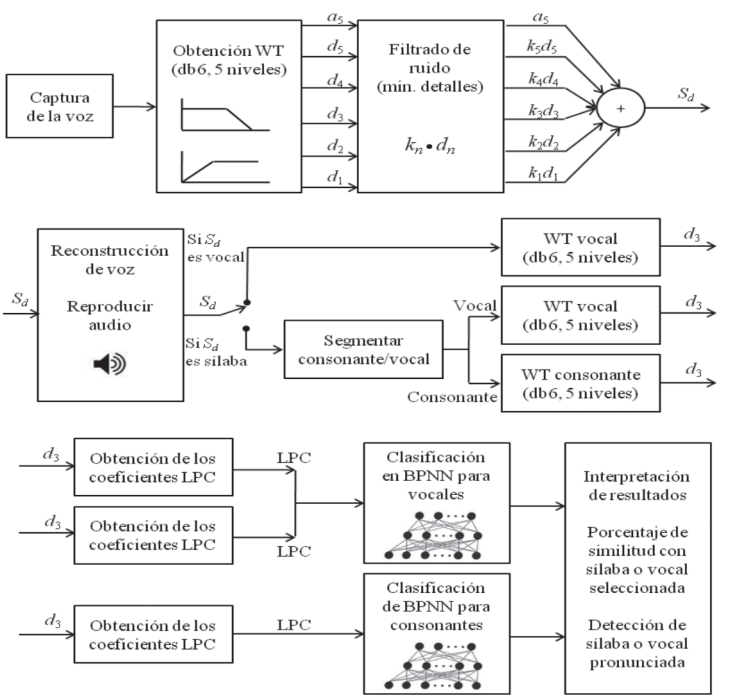

Figura 2. Modelo final para reconocimiento de vocales y sílabas.

Los LPC son interpretados por la BPNN para la identificación de la vocal, indicándose el porcentaje de parentesco con la vocal ingresada. Como ejemplo, se muestra el procesamiento de la vocal "a", como se muestra en la Figura 3.

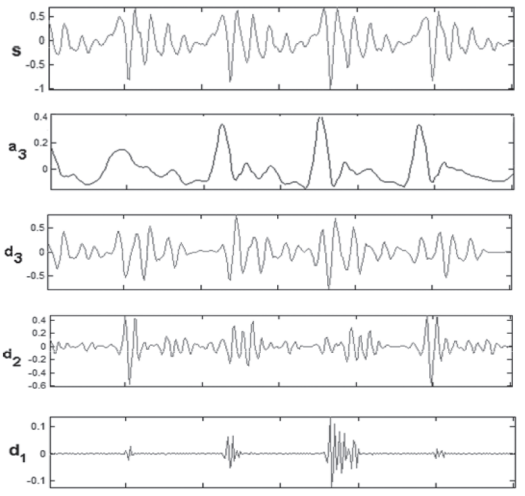

Figura 3. Descomposición de la vocal "a" en 4 niveles de resolución.

Cuando d_3 proviene de una sílaba se realiza un proceso de segmentación o separación entre la consonante y la vocal que componen la sílaba [26], de forma de realizar el procesamiento de estos dos componentes separadamente y así facilitar su identificación. La vocal componente de la sílaba se procesa de igual modo.

En el caso de la consonante de la sílaba, también se procesa a través de la aplicación de la WT y posterior extracción de los parámetros LPC, los cuales son reconocidos a través de una red neuronal especializada para el reconocimiento de consonantes.

Para el caso de una sílaba, una vez reconocida su consonante y su vocal se procede a la identificación propiamente tal. Para comprender la nomenclatura utilizada, es necesario aclarar dos conceptos; ejemplos y muestras. Un ejemplo se constituye por un conjunto muestras que lo caracterizan, es decir, un ejemplo es la vocal "a", una sola pronunciación, y está constituida por un determinado número de muestras.

Debido a la naturaleza compleja de la voz humana, donde se puede decir que existen tantas voces como personas con habla, la cantidad de individuos que deben participar para independizar a un sistema artificial del hablante, es aún una incógnita; se posee tan solo la intuición de que mientras más variado sea el universo de ejemplos, mejores resultados se obtendrán. Todo este proceso se desarrolla únicamente para el idioma español.

También para un adecuado funcionamiento de una ANN existen reglas básicas de la cantidad de ejemplos a utilizar en las fases de entrenamiento, de pruebas y validación, y que se encuentra directamente relacionada con la cantidad de muestras que componen cada ejemplo.

La propuesta desarrollada contempló, en primera instancia, el ingreso directo a la ANN de ejemplos d_3 con sus respectivas muestras. Ello trajo como inconveniente la necesidad de una gran cantidad de datos (ejemplos) para su entrenamiento. A continuación se muestran las implicancias de esta metodología.

Para la recolección de ejemplos, se emplea un programa computacional que recopila cada sílaba o vocal pronunciada y en forma automática la adapta y la nombra para ser un ejemplo válido que ingrese a la base de datos para el entrenamiento. La frecuencia de muestreo empleada es de 8000 Hz, considerando el ancho de banda de 4000 Hz usado en telefonía, que contiene la información característica de la voz humana. Luego, de manera experimental se determina que la vocal que posee el mayor pitch, o período fundamental, es la vocal "a". A partir de los ejemplos recolectados de todos los hablantes grabados, se determina que con la cantidad de 80 muestras, es decir 10 ms, bastan para determinar el pitch. Esta cantidad es la mínima con la que se podría construir la vocal "a" y por tanto puede usarse para ingresar a la red neuronal directamente.

Se entrena una BPNN sencilla con ejemplos de 80 muestras para identificar las 5 vocales. Para ello, se debe disponer de un total de 4350 ejemplos, es decir 870 por cada vocal, lo que demandaría tiempo y recursos excesivos. Se hace necesario entonces, reducir la cantidad de ejemplos, y la variable a manejar es la cantidad de muestras componentes de cada uno y que determinan la cantidad de neuronas de entrada (*Ne*) en la ecuación 2. En el caso de las sílabas, se utilizan redes distintas que las usadas

para las vocales, agrupándolas en conjuntos de 4 o 5 por red.

Preparación de ejemplos para el entrenamiento
La metodología propuesta para lograr reducir la cantidad de muestras de cada ejemplo consiste en utilizar la WT y los coeficientes LPC. Esta idea se fundamenta en la capacidad de la WT para reducir el ruido y su eficiencia para representar la señal con menos muestras. Por otra parte, la utilidad de los coeficientes LPC para representar paramétricamente una señal de voz, los cuales pueden reemplazar a un gran número de muestras por un conjunto pequeño de parámetros llamados coeficientes LPC. Mediante WT se realiza un filtrado de ruido a la totalidad de ejemplos para eliminar la información no deseada para luego realizar una descomposición de 5 niveles mediante la wavelet madre Daubechies 6, que fue la que entregó mejores resultados.

La Figura 3 muestra la descomposición wavelet de la vocal "a" en 5 niveles de resolución. Se observa que una de las ondas de menor resolución (con sólo 38 muestras), correspondiente al detalle d_3, posee una gran similitud con la señal original S (de 300 muestras). Si se selecciona el detalle d_3, con sólo 38 muestras, se necesitarían 2250 ejemplos, 450 por cada vocal, para el entrenamiento, es decir, se reduce el universo al 51% de la cantidad de 4350 ejemplos mencionada anteriormente.

El uso de los coeficientes LPC para el reconocimiento de la voz, es una de las técnicas más utilizadas y con muy buenos resultados. La cantidad necesaria de coeficientes para lograr una adecuada representación de la señal sometida a análisis se relaciona de forma directa con la frecuencia de muestreo y se considera, mediante trabajos experimentales, que con 13 coeficientes basta para una señal muestreada a 8 kHz. Si se seleccionasen 12 coeficientes LPC, descartando el décimo tercero, por ser siempre 1, debido a la normalización con respecto al último coeficiente, para entrenar las redes, según (2) se necesitarían 950 ejemplos (190 por cada vocal).

Si el detalle d_3 es una representación de buena calidad y de baja resolución de la señal original, puede considerarse que la frecuencia de muestreo baja en un 77%, entonces se podría utilizar una menor cantidad de coeficientes LPC.

En el proceso experimental se determinó que el límite inferior de coeficiente con el cual los resultados del entrenamiento se mantienen con mínimas variaciones es de 7 y que con una cantidad menor los resultados caen drásticamente. Al utilizar 6 coeficientes (el séptimo es 1 por la normalización) según (2) se necesitan sólo 650 ejemplos (140 por vocal) para entrenar una red. Así se redujo al 15% la cantidad de ejemplos mínimos necesarios para llevar a cabo el entrenamiento de la red.

VALIDACIÓN DEL MODELO

Para validar una ANN se poseen indicadores estadísticos y de error que muestran el comportamiento real de la red. Tres de estos indicadores son: el error cuadrático medio (RMS, Root Mean Square Error), la desviación estándar residual (RSD, Residual Standard Deviation) y el índice de adecuación (AI, Adequation Index). Teniéndose que, idealmente, estos valores llegan a cero en el caso de RMS y RSD, y a uno para AI [27].

Diseño del experimento
El experimento desarrollado consiste en generar variadas ejemplos a partir de diversos hablantes de habla hispana de Chile (país donde se ubica el estudio). Se muestran los resultados para las 5 vocales y 4 fonemas característicos: "ya", "te", "pi", "bu".

Resultados del entrenamiento y pruebas
Durante el proceso experimental se entrenaron y probaron diferentes configuraciones de redes, respetando las limitaciones expresadas en (2), utilizando distintos tipos de entradas, así como funciones de activación y algoritmos de entrenamiento. De todas ellas, las que mejores resultados entregaron, en términos de menor error de entrenamiento y mayor simplicidad, son las que a continuación se detallan en las tablas siguientes en términos de sus arquitecturas y respectivas configuraciones topológicas.

Resultados BPNN para reconocimiento de vocales
A continuación, en la Tabla 1, se presentan las características de la BPNN utilizada para la caracterización de las vocales.

Luego, los resultados de las pruebas de la red de vocales se presentan en la Tabla 2 y para las sílabas, en la Tabla 3.

Tabla 1. Características de las vocales.

Red de vocales "a", "e", "i", "o", "u"	
Ejemplos de entrada a la red	LPC del datalle "d3"
Cantidad de muestras por ejemplo	6 (6 de 7 coef. LPC)
Tipo de red	Backpropagation
Neuronas de entradas	6
Neuronas de capa oculta	5
Neuronas en capa de salida	5
Función de transferencia capa oculta	Tansing (sigmoidal)
Función de transferencia capa salida	Purelin (lineal acotada)
Algoritmo de entrenamiento	Trainlm

Tabla 2. Resultados de pruebas de red de vocales.

	Aciertos en total por cada vocal en 36 ejemplos (ideal: 100%)					Promedio resultados 36 ejemplos x vocal (ideal: 1)				
	a	e	i	o	u	a	e	i	o	u
a	34	0	2	0	0	0,92	0,06	0,08	0,11	0,09
	94%	0%	6%	0%	0%					
e	0	32	2	2	4	0,04	0,71	0,16	0,13	0,2
	0%	89%	6%	6%	11%					
i	2	1	24	1	8	0,11	0,19	0,52	0,13	0,27
	6%	3%	67%	3%	22%					
o	0	2	4	32	0	0,05	0,14	0,15	0,77	0,13
	0%	6%	11%	89%	0%					
u	0	1	4	1	24	0,06	0,16	0,23	0,13	0,54
	0%	3%	11%	3%	67%					
Tot	36	36	36	36	36	--	--	--	--	--
Pro	81% de acierto					0,69 de similitud				

Observación: lo que está sombreado son los aciertos (equivalencia entre la vocal ingresada y la pronunciada), y lo que está fuera de las diagonales son los parentescos de la vocal ingresada con las otras vocales.

Tabla 3. Resultados de pruebas de red de sílabas.

	Aciertos en total por cada sílaba en 36 muestras (ideal 100%)				Promedio resultados 36 muestras por sílaba (ideal 1).			
	ya	te	pi	bu	ya	te	pi	Bu
ya	23	9	3	2	0,57	0,37	0,13	0,08
	64%	25%	8%	6%				
te	8	27	1	0	0,24	0,63	0,08	0,03
	22%	75%	3%	0%				
pi	2	0	25	6	0,17	0,05	0,58	0,33
	6%	0%	69%	17%				
bu	3	0	7	28	0,14	0,04	0,29	0,61
	8%	0%	19%	78%				
total	36	36	36	36	--	--	--	--
Pro	72% de asertividad				0,60 de similitud			

Resultados BPNN para reconocimiento de sílabas

La red presenta la misma arquitectura y entrenamiento que la red de vocales, excepto el número de salidas, que son sólo 4, debido a la cantidad de sílabas a reconocer.

Resultados de validaciones

La validación cuantitativa permite, a través de los indicadores de error, medir la convergencia del sistema. A continuación se presentan estos índices para ambas redes: de vocales y sílabas, mediante las Tablas 4 y 5.

ANÁLISIS DE RESULTADOS

Empleando la totalidad de los ejemplos para entrenamiento (130 por cada vocal), se obtuvieron los coeficientes LPC que se ingresan a la red de vocales y se alcanzó un 81% de asertividad, según se aprecia en la Tabla 2.

El nivel de similitud obtenido para el reconocimiento de vocales es de 0,69. Siendo la vocal "a" la más fácil de reconocer con 94% de acierto y 0,92 de similitud. En cuanto a los índices de error, se tiene que el RMS promedio es de 0,348. El menor error lo entrega la vocal /a/ con RMS = 0,146. Para el RSD, se aprecia que el valor promedio es de 0,275, y nuevamente la vocal /a/ es la que presenta mejores resultados. El AI es 0,432 en promedio para la red de vocales, presentado en la Tabla 4.

Tabla 4. Indicadores de validación (RMS, RSD e IA) para red de vocales.

Vocal	RMS	RSD	IA
a	0,146	0,133	0,443
e	0,354	0,276	0,427
i	0,401	0,302	0,374
o	0,388	0,314	0,442
u	0,452	0,35	0,472
Promedio	**0,348**	**0,275**	**0,432**

Para la red de sílabas, se utilizó un conjunto de 520 ejemplos. Al observar la Tabla 3 se tiene que la asertividad fue de 72% y el nivel de similitud de 0,60. Durante la validación se obtuvieron los índices RMS y RSD, que se muestran en la Tabla 5, con un promedio de 0,552 y 0,375, respectivamente. El AI tuvo un valor de 0,42.

Tabla 5. Indicadores de validación (RMS, RSD e IA) para red de sílabas.

Sílaba	RMS	RSD	IA
ya	0,589	0,35	0,426
te	0,488	0,341	0,427
pi	0,565	0,378	0,407
bu	0,567	0,432	0,438
Promedio	**0,552**	**0,375**	**0,424**

CONCLUSIONES

A partir del análisis de los resultados, es posible afirmar que el modelo propuesto para el reconocimiento de sílabas y vocales, implementado a través de un software apoyado en las teorías de WT, coeficientes LPC y BPNN, ofrece interesantes perspectivas para constituirse en una herramienta automática y amigable para complementar el aprendizaje del habla para personas que padecen problemas de audición. Este modelo es extensible a palabras, utilizando la segmentación de las mismas.

Respecto de los resultados cuantitativos, se destaca la gran ventaja de emplear la WT, disminuyéndose la cantidad de muestras de la señal. La incorporación de los parámetros LPC logra bajarla a un 2,3%, ya que su utilización aplicada al detalle d_3 de la WT de la señal, permite en forma mejorar y optimizar la capacidad de entrenamiento de las redes neuronales, las cuales fueron entrenadas con ejemplos de baja complejidad permitiendo, de esta manera, utilizarlo en línea.

El modelo desarrollado es implementado en una aplicación con una interfaz gráfica amigable a nivel de prototipo, la cual será descrita en un próximo trabajo, posibilitando una utilidad práctica real.

Se espera que este sistema sea evaluado a futuro por un experto en lenguaje, de manera de tener un respaldo de esta disciplina, sobre los resultados numéricos presentados. También se tiene la perspectiva de incorporar nuevas sílabas y palabras usando segmentación, así como también realizar pruebas con personas con dificultades de audición a través de la interfaz presentada en la propuesta.

AGRADECIMIENTOS

Este trabajo ha contado con el apoyo de la Universidad de Santiago de Chile, a través del Convenio de aportes basales por desempeño del proyecto MECESUP USA 1298.

REFERENCIAS

[1] NICHCY. “La sordera y la pérdida de la capacidad auditiva”. Academy for Educational Development. 2010. Fecha de Consulta: 20 de enero de 2013. URL: http://www.sfusd.edu/es/assets/sfusd-staff/programs/files/special-education/ sordera-SP.pdf

[2] D. Reccasens. “Fonètica i Fonologia”. Biblioteca Universitaria 18. Enciclopedia Catalana. España. 1993.

[3] A. Paul, D. Das and M. Kamal. “Bangla Speech Recognition System Using LPC and ANN”. Seventh International Conference on Advances in Pattern Recognition, ICAPR, pp. 171-174. 2009.

[4] C. San Martín y R. Carrillo. “Implementación de un reconocedor de palabras aisladas dependiente del locutor”. Revista Facultad de Ingeniería-Universidad de Tarapacá. Vol. 12 Nº 1, pp. 9-14. ISSN: 0718-1337. 2004.

[5] C. Fook. “A review: Malay speech recognition and audio visual speech recognition”. International Conference on Biomedical Engineering (ICoBE), pp. 479-484. 2012.

[6] M. Kaynak, Q. Zhi, A. Cheok, K. Sengupta and K. Chung. “Audio-visual modeling for bimodal speech recognition”. IEEE Systems, Man, and Cybernetics. Vol. 1, pp. 181-186. 2001.

[7] H. Sheikhzadeh. “Waveform-based speech recognition using hidden filter models: parameter selection and sensitivity to power normalization”. IEEE Transactions on Speech and Audio Processing. Vol. 2 Nº 1, pp. 80-89. 1994.

[8] R. Gandhiraj and P. Sathidevi. “Auditory-Based Wavelet Packet Filterbank for Speech Recognition Using Neural Network”. International Conference on Advanced Computing and Comm. ADCOM, pp. 666-673. 2007.

[9] C. Medina, A. Alcaim and J. Apolinario. “Wavelet denoising of speech using neural networks for threshold selection”. Electronics Letters. Vol. 39 Nº 25, pp. 1869-1871. 2003.

[10] K.Daqrouq, A.R. Al-Qawasmi, K.Y. Al Azzawi and T. Abu Hilal. “Discrete Wavelet

Transform & Linear Prediction Coding Based Method for Speech Recognition via Neural Network". INTECH Discrete Wavelet Transforms-Biomedical Applications. Vol. 12, pp. 117-132. 2011.

[11] Khaled Daqrouq, Ali Morfeq, Mohammad Ajour and Abdulhameed Alkhateeb. "Wavelet LPC With Neural Network for Speaker Identification System". WSEAS Transactions on Signal Processing. Vol. 9, pp. 216-226. 2013.

[12] G. Bonilla-Enriquez and S. Caballero-Morales. "Communication interface for mexican spanish dysarthric speakers. Redalyc, Scientific Information System". Network of Scientific Journals from Latin America, the Caribbean, Spain and Portugal. Vol. 22 Nº 1, pp. 98-105. 2012.

[13] S. Domínguez, E. Salama and J. García-Bermejo. "Arisco, un robot social con capacidad de interacción, motivación y aprendizaje". Revista Iberoamericana de Automática e Informática Industrial. Vol. 5 Nº 2, pp. 69-78. 2008.

[14] A. Jardón, A. Giménez, R. Correal, S. Martínez and C. Balaguers. "Asibot: Robot portátil de asistencia a discapacitados. Concepto, arquitectura de control y evaluación clínica". Revista Iberoamericana de Automática e Informática Industrial. Vol. 5 Nº 2, pp. 48-59. 2008.

[15] L. Salhi, M. Talbi and A. Cherif. "Voice Disorders Identification Using Hybrid Approach: Wavelet Analysis and Multilayer Neural Networks". International Journal of Computer, Information, Systems and Control Engineering. Vol. 2, pp. 193-202. 2008.

[16] N. Balsero, D. Botero, J. Zuluaga y C. Parra. "Interacción hombre-máquina usando gestos manuales en texto real". Network of Scientific Journals from Latin America, the Caribbean, Spain and Portugal. Vol. 9 Nº 2, pp. 1-13. 2005.

[17] Sigurdsson, Brandt and Lehn-Schiler. "Mel Frequency Cepstral Coefficients. An Evaluation of Robustness of MP3 Encoded Music". Proceedings of the Seventh International Conference on Music Information Retrieval (ISMIR). 2006.

[18] E. San Juan, F. Watkins y H. Kaschel. "Sistema de ayuda visual para apoyar aprendizaje de fonemas españoles". Revista de Ingeniería Electrónica, Automática y Comunicaciones. Vol. 34 Nº 1, pp. 87-99. 2013.

[19] M. Jamett. "Feedforward Convergence and Stability Analysis from a Set Perspective: State Estimation Approach". Tesis para optar al grado de doctor. Universidad de Santiago de Chile. Santiago, Chile. 2004.

[20] M. Jamett and G. Acuña. "Comparative assessment of interval and affine arithmetic in neural network state prediction". Lecture Notes in Computer Science. Vol. 3497, pp. 448-453. 2005.

[21] P. Faúndez. "Procesamiento digital de señales acústicas utilizando wavelets". Memoria de titulación de Ingeniería Acústica. Universidad Austral de Chile. Valdivia, Chile. 1999.

[22] J. Walter. "A primer on wavelets and their scientific applications". University of Winconsin-Eau Claire, Hall/CRC. 1999. URL: http://dsp-book.narod.ru/WaMW.pdf. Fecha de Consulta: 25 de abril de 2013.

[23] T. Qian, Y. Xu and M. Vai. "Wavelet analysis and applications". Editorial Birkhäuser. Berlín, Alemania. 2007.

[24] T. Parsons. "Voice and speech processing". McGraw-Hill Series in Electrical and Computer Engineering. New York, USA. 1987.

[25] M. Faúndez. "Tratamiento Digital de Voz e Imagen". Editorial Marcombo. México. 2000.

[26] E. San Juan, H. Kaschel, F. Watkins y P. López. "Segmentación de sílabas y fonemas". XIII Congreso Internacional de Telecomunicaciones SENACITEL. Valdivia, Chile. 2008.

[27] M. Spiegel and J. Schiller. "Srinivasan Schaum's outline of probability and statistics". 2° ed. Editorial McGraw-Hill. México, DF. 2000.