Universidad de Alcalá

XXVIII CONVOCATORIA DE PREMIOS
"INGENIEROS DE TELECOMUNICACIÓN"
Proyecto Fin de Carrera

CLASIFICACIÓN AUTOMÁTICA DE SEÑALES SONORAS APLICADA A LA MEJORA DE LA INTELIGIBILIDAD DE LA VOZ EN AUDÍFONOS DIGITALES

Lorena Álvarez Pérez

Dept. Teoría de la Señal y Comunicaciones
Universidad de Alcalá

Índice

1. Descripción del trabajo realizado **2**
1.1. Origen . . . 2
1.2. Objetivos . . . 3
1.3. Desarrollo . . . 4
1.3.1. Creación de la base de datos . . . 5
1.3.2. Estudio y extracción de características para la clasificación . . . 5
1.3.3. Estudios de algoritmos de clasificación sencillos . . . 6
1.3.4. Estudios de las Redes Neuronales Artificiales como algoritmos de clasificación 6
1.4. Conclusiones . . . 7

2. Originalidad **8**

3. Resultados **9**

4. Aplicabilidad **12**

Anexo I. Financiación **14**

Anexo II. Publicaciones **15**

1. Descripción del trabajo realizado

1.1. Origen

Hoy en día, aproximadamente el 13 % de la población de los países desarrollados sufre pérdidas de audición considerables, cuyas consecuencias negativas podrían ser compensadas haciendo uso de alguna clase de audífono [1]. Estas pérdidas no sólo afectan de forma significativa a la comunicación oral, sino que además, deterioran la calidad de vida de las personas con deficiencias auditivas. Esto ha motivado a científicos e ingenieros a desarrollar audífonos digitales más avanzados, que sean más confortables y manejables especialmente para las personas mayores, dado que la pérdida de audición es una de las condiciones crónicas más prevalentes en la tercera edad.

La utilización de la tecnología digital para el desarrollo de **audífonos digitales** es un tema de investigación abierto, donde, por un lado, hay muchos problemas sin resolver, y, por otro, algunas de las soluciones propuestas no son enteramente satisfactorias. Los audífonos digitales modernos proporcionan varios "programas", cada uno de los cuales está recomendado para una situación acústica diferente, de modo que puedan modificar su respuesta según las características acústicas del entorno buscando siempre optimizar la percepción del lenguaje oral. Estos programas se activan por medio de pulsadores o controles remotos. Sin embargo, no son muy adecuados para personas mayores, ya que requieren que el anciano detecte el entorno y seleccione el programa. Sería mucho más práctico para el usuario que el propio audífono fuese capaz de detectar de forma automática el entorno sonoro en que se encuentra el paciente y activar el programa mejor adaptado a tal entorno. Mejoraría ello la percepción de la voz y el nivel de confort.

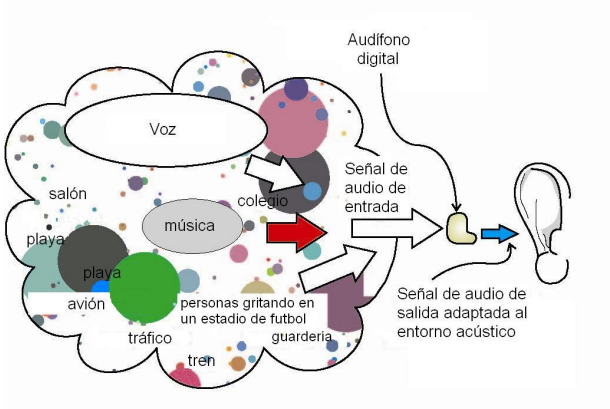

Figura 1: Representación de la variedad de sonidos a los que los usuarios de los audífonos se suelen enfrentar a diario (voz, música, tráfico, gente gritando en un estadio de fútbol, etc.).

Con respecto a la variedad de entornos acústicos a los que el paciente tiene que hacer frente en su vida normal (ver Fig. 1), se ha demostrado que los usuarios, generalmente, prefieren tener un número de esquemas de amplificación adaptados a diferentes condiciones auditivas [2,3]. Aunque el ambiente acústico tiene muchos aspectos, se ha demostrado que las situaciones diarias se pueden agrupar en cuatro grupos principales que cubren la práctica totalidad de las situaciones auditivas y que permiten a un usuario medio llevar una vida "normal": conversación en ambiente tranquilo, conversación en ambiente ruidoso, ruido y música. Como la mayor parte de los usuarios suelen dar mayor importancia a poder comprender la voz, en este proyecto se optó por agrupar el ruido y la música (tanto vocal como instrumental) dentro de una única clase denominada no-voz.

A este respecto, la etapa más importante en un audífono es la **discriminación entre voz y no voz**, conteniendo la primera clase tanto conversaciones en "ambientes tranquilos" como conversaciones en "ambientes ruidosos". Para explicar la importancia de la clasificación entre voz y no-voz, resulta instructivo considerar las dos siguientes situaciones:

- Situación 1: Imaginemos que el usuario se encuentra, por ejemplo, en una conferencia. Si el sistema clasifica la voz como "no voz", el audífono probablemente decidirá que **no** es necesario amplificar tal señal, y por tanto, reducirá la ganancia. La consecuencia sería que la persona con deficiencia auditiva perderá toda la información contenida en el fragmento de voz. Esto degrada la utilidad de los audífonos.

- Situación 2: El paciente se encuentra ahora en un entorno ruidoso, por ejemplo, en un atasco de tráfico. Si el sistema clasifica el sonido como voz, entonces el audífono decidirá que sí es necesario amplificar la señal. La consecuencia inmediata es que el usuario escuchará de repente un ruido amplificado, irritante y molesto. Esto degrada la comodidad.

Estas dos situaciones ilustran el hecho de que la tarea más crucial es la discriminación entre voz y no voz en un audífono. Una vez realizada esta primera etapa de clasificación, si la señal se clasificó como voz, se puede realizar una segunda **clasificación entre conversación en ambiente tranquilo ("voz limpia") y conversación en ambiente ruidoso ("voz con ruido")**.

Tan importante como esto resulta el hecho de que la mayoría de los audífonos digitales basados en DSPs (Digital Signal Processors), que hay en el mercado presentan **importantes restricciones en términos de complejidad computacional, memoria y batería**. Estas restricciones son debidas principalmente al reducido tamaño del audífono (especialmente para los modelos ITC -In The Canal- y CIC -Completely In the Canal-), el cual, tal y como se ilustra en la figura 2, contiene no sólo los dispositivos electrónicos sino también la batería para alimentarlos. Si además, se tiene en cuenta que los algoritmos necesarios para compensar las pérdidas acústicas requieren una parte considerable de la complejidad computacional, estamos obligados a usar **algoritmos de clasificación de baja complejidad**.

Estas restricciones pueden entenderse mejor si tomamos como ejemplo el DSP que se utilizará para la implementación real del audífono. Este procesador trabaja a una frecuencia de reloj de 2 MHz, lo que, en el mejor de los casos, implica una capacidad computacional de alrededor de 2 MIPS. Teniendo en cuenta que se trabaja con una frecuencia de muestreo de 16 kHz, se dispone en total de 125 operaciones por muestra. Estas 125 operaciones, que deben ser entendidas como instrucciones de ensamblador, tienen que ser suficientes para realizar todo el procesado básico de compensación de las pérdidas auditivas (amplificación, compresión por tramos, etc.), así como para implementar el algoritmo de clasificación. A la vista de estos números queda clara la enorme complejidad del problema planteado, así como la acuciante necesidad de reducir la complejidad computacional de todos los algoritmos necesarios para la clasificación de los sonidos.

1.2. Objetivos

El **objetivo final** del proyecto fue **diseñar un sistema de clasificación robusto** e **implementable en un audífono digital** que, de forma automática, fuese capaz, por sí solo, de reconocer en qué entorno acústico se encuentra el usuario y adaptarse a dicho entorno activando el mejor programa para tal situación. Se ha buscado con ello mejorar la inteligibilidad de la palabra y el consiguiente confort del paciente. Para alcanzar el objetivo global, se plantearon los siguientes **objetivos parciales**:

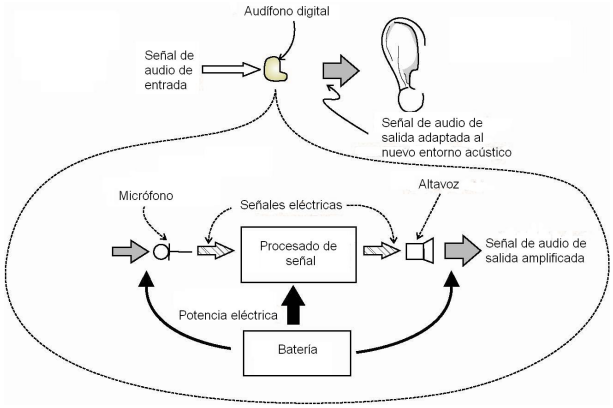

Figura 2: Diagrama simplificado de la estructura de un audífono típico. El micrófono transforma el sonido de entrada en una señal eléctrica que puede ser utilizada por el bloque de procesado de señal para compensar las pérdidas que el paciente padece. El altavoz convierte esta señal modificada y amplificada en un sonido adaptado a las deficiencias del paciente.

- Análisis y caracterización de la información presente en la señales de audio viable de ser utilizada en un sistema de reconocimiento y clasificación automático.
- Estudio de diferentes técnicas de selección de las características más relevantes para la clasificación de señales de audio.
- Estudio de distintos algoritmos para la clasificación automática de sonidos. En particular, se estudiaron los siguientes algoritmos: k-NN, discriminante lineal de Fisher y las Redes Neuronales Artificiales (RNAs).
- Desarrollo de un sistema que haga uso del clasificador (o clasificadores) escogido(s) para la correcta clasificación e identificación de entornos sonoros.

Es resañable decir que, si bien, estos eran los objetivos iniciales del proyecto final de carrera, **su consecución ha dado lugar a otros nuevos objetivos**. Así, actualmente se sigue trabajando a partir de todo lo diseñado en este proyecto con un nuevo y ambicioso objetivo principal: **la elaboración de la tesis doctoral de la proyectista cuyo objetivo final es diseñar una prototipo funcional de audífono digital que incorpore técnicas avanzadas de procesado de la señal y que sea capaz de adaptarse de forma automática al entorno sonoro en el que se encuentre el paciente**.

1.3. Desarrollo

En una primera **fase exploratoria**, el trabajo se centró en realizar una exhaustiva revisión bibliográfica de los distintos algoritmos propuestos para la clasificación automática de señales, así como de las características más apropiadas para esta tarea. Con toda esta información se pudo construir el marco teórico, que integraba todas las teorías, estudios y antecedentes relacionados con el problema a investigar, y que sustentó todo el trabajo posterior.

La **hipótesis de partida** es que es posible desarrollar algoritmos de bajo coste computacional (una de las limitaciones de diseño mencionadas) capaces de clasificar de forma automática el entorno en el que se encuentre el paciente, y, a partir de esta información, activar las técnicas

necesarias para aumentar la calidad subjetiva del audio que escucha el paciente. Tal y como se ha comentado, si la señal detectada es de voz, el objetivo será incrementar la inteligibilidad percibida. Por otra parte, si es ruido, se evitará amplificarlo en exceso, a fin de no crear molestias innecesarias al paciente.

Tomando como punto de partida esta hipótesis, los pasos necesarios para alcanzar el objetivo de la clasificación automática de entornos sonoros en audífonos digitales fueron los que se describen a continuación.

1.3.1. Creación de la base de datos

Fue necesario diseñar una base de datos nueva que sirviera de base para la mayor parte de los experimentos realizados en el proyecto y que ampliase asimismo la base de datos de sonidos, utilizada hasta entonces, para clasificación voz/música proporcionada por Daniel W. Ellis. Esta nueva base de datos contenía un total de 2936 señales de audio divididas en voz, voz intercalada con ruidos de fondo (con relaciones señal a ruido de 20 dB, 10 dB y 0 dB), música (vocal e instrumental) y ruido.

Para crear los ficheros de voz intercalada con ruido de fondo se seleccionaron los ficheros que únicamente contenían voz de la base de datos de Daniel W. Ellis y se mezclaron con los distintos tipos de ruido tomados de múltiples fuentes. Tanto los ficheros de voz como los ficheros que contenían ruido se fragmentaron en ficheros de duración 2,5 segundos. En el caso de los ficheros de ruido fue necesario, además, remuestrearlos con una frecuencia de muestreo de 16000 Hz y 16 bits por muestra.

Es interesante recalcar en este punto la **novedad** que implica la creación de esta base de datos, al no existir en la actualidad ninguna base de datos pública específica para el sonido de audífonos digitales.

1.3.2. Estudio y extracción de características para la clasificación

El proceso de extracción de características es de gran importancia para alcanzar el resultado final del clasificador. Es necesario seleccionar una serie de parámetros de la señal de audio de entrada que sean suficientemente representativos de cara a la futura clasificación. Dependiendo del problema en particular, estas características podrán variar, y así no se utilizarán los mismos parámetros para diferenciar voz de música, rock de jazz o blues de country.

La mayor parte de los parámetros utilizados en las implementaciones encontradas en la literatura se pueden dividir en tres tipos, dependiendo de si están relacionados con el timbre, el ritmo o el *pitch* de la señal. Tanto para el caso de discriminación voz/no-voz o voz/música, se suele aceptar que basta con utilizar los parámetros basados en el timbre para obtener muy buenos resultados, por lo que, en este trabajo, la mayor parte de los parámetros utilizados están basados en el timbre.

En este proyecto se han extraído un total de catorce características de cada señal de audio de entrada, especialmente adaptadas al problema de la clasificación automática de sonidos en audífonos digitales. Estas características han sido: Spectral Centroid, Spectral Roll-Off, Spectral Flux, Zero Crossing Rate Ratio (ZCR), High Zero Crossing Rate Ratio (HZCRR), Short Time Energy (STE), Low Short-Time Energy Ratio (LSTER), Mel-Frequency Cepstral Coefficients (MFCCs), Voice2White, Percentage of Low Energy Frames (LEF), Loudness, Spectral Flatness Measure (SFM) y Bandwidth.

1.3.3. Estudios de algoritmos de clasificación sencillos

Debido a las ya mencionadas limitaciones que presentan los audífonos digitales basados en DSPs, los primeros experimentos en clasificación voz/no-voz se realizaron con dos algoritmos de clasificación sencillos. Estos dos algoritmos son el discriminante lineal de Fisher y el algoritmo k-NN (k-Nearest Neighbours), este último un clasificador muy popular en las tareas de audio. El planteamiento de ambos métodos es simple e intuitivo.

El discriminante lineal de Fisher es un método estadístico de clasificación que se caracteriza por tener una fuerte base teórica, mientras que el algoritmo k-NN es una técnica de discriminación no paramétrica que se engloba dentro de los algoritmos de clasificación por vecindad.

Para ambos métodos de clasificación, se realizaron experimentos utilizando tanto las características por separado como distintas combinaciones de las mismas. Asimismo, se intentó buscar la combinación de características y clasificadores que mejores resultados diese en términos de porcentaje de clasificación correcta.
Estos primeros experimentos sirvieron para saber qué características se comportaban mejor en la discriminación voz/no-voz.

1.3.4. Estudios de las Redes Neuronales Artificiales como algoritmos de clasificación

Las Redes Neuronales Artificiales (RNAs) son sistemas inspirados en la funcionalidad de las neuronas biológicas, y han resultado ser especialmente aptas para modelar y efectuar predicciones en sistemas muy complejos [4]. Con objeto de mejorar los resultados obtenidos, se aplicaron las RNAs tanto al problema de discriminación voz/no-voz como al problema de discriminación voz limpia/voz con ruido. Se realizaron pruebas con tres algoritmos de entrenamiento distintos como son el sencillo descenso por gradiente, con función objetivo error cuadrático medio y entropía, y dos más complejos, como el método de Levenberg-Marquardt y el método Levenberg-Marquardt basado en técnicas de regularización bayesiana. Con cada uno de los métodos de entrenamiento, se realizaron experimentos tanto para cada característica por separado como para distintos conjuntos de características, eligiendo aquellas características que de forma aislada mejor separabilidad proporcionaban. Con el objetivo de comparar los resultados obtenidos, los mismos experimentos se repitieron para el discriminante lineal de Fisher y el algoritmo k-NN.

Un handicap a la hora de utilizar las RNAs como clasificadores en audífonos digitales basados en DSPs, podría ser, a priori, su alta complejidad computacional, que entre otros temas, está relacionada con el tamaño de la red. Sin embargo, merece la pena su implementación porque, como se demostró en [1], las RNAs son capaces de proporcionar muy buenos resultados en términos de porcentaje de acierto en comparación con otros algoritmos como el discriminante lineal de Fisher, el clasificador de mínima distancia, el algoritmo k-NN o el clasificador Bayesiano. La única desventaja de las RNAs, como se mencionó, es su alta complejidad computacional. Esta complejidad depende del número de pesos que necesitan adaptarse, y en consecuencia, del número de neuronas que componen la red. Por tanto, interesa reducir el número de neuronas en la capa oculta. Es por esta razón que en los experimentos del proyecto, el número de neuronas en la capa oculta se probó a variar de 1 a 30, eligiendo el mejor valor en términos de error de validación. Asimismo, se realizaron experimentos aplicando la técnica VSL (Variable Structure Learning) para el dimensionado automático de las RNAs. Esta técnica, que permite minimizar la estructura de la red neuronal manteniendo una tasa de acierto razonable, fue evaluada bajo los distintos métodos de entrenamiento, aunque se hizo de forma más exhaustiva en la tarea clasifiacatoria voz/no-voz con el método de descenso por gradiente y función objetivo error cuadrático medio,

por ser el algoritmo de menor coste computacional. En el caso de discriminación voz limpia/voz con ruido, esta técnica se aplicó a distintos conjuntos de características con el método descenso por gradiente y función objetivo el error cuadrático medio.

1.4. Conclusiones

El trabajo realizado en este proyecto se ha centrado principalmente en el desarrollo de un clasificador de bajo coste computacional para la detección automática del entorno sonoro en el que se encuentra el paciente.
Tras el estudio teórico, el desarrollo de distintos clasificadores y simulaciones, la aplicación de diferentes técnicas para el dimensionado automático de las RNAs y la búsqueda de distintas combinaciones tanto de características como de clasificadores que mejorasen los resultados obtenidos, las principales conclusiones del proyecto (publicadas en las 14 referencias del anexo II) son:

1. **Discriminación voz/no-voz**

   - En una evaluación para una **base de datos reducida con** 546 **señales de audio**, y para dos métodos de clasificación sencillos como son el **discriminante lineal de Fisher y el algoritmo *k*-NN**, utilizando una *combinación de cinco clasificadores* cada uno de ellos con **distintas características a la entrada**, se ha conseguido alcanzar una **tasa de acierto del** 97,0 %. Las cifras obtenidas demuestran que existen ciertas características que combinadas de forma adecuada proporcionan buenos resultados para esta tarea.
   - En base a obtener resultados más generalizados, se evaluó la eficiencia de las **RNAs** para la discriminación voz/no-voz. Para estos experimentos se utilizó la **base de datos diseñada** para el proyecto y que constaba de 2936 **señales de audio**. Los resultados obtenidos permitieron concluir que el **mejor resultado** se obtenía para el **método Levenberg-Marquardt con técnicas de regularización bayesiana** cuando se utilizaban todas las características, logrando un **porcentaje de acierto del** 96,6 %. Los mismos experimentos se repitieron para el discriminante lineal de Fisher y el algoritmo k-NN, obteniendo siempre peores resultados.

2. **Discriminación voz limpia/voz con ruido**

   Para esta tarea de clasificación, se utilizó la base de datos diseñada específicamente para el proyecto. El **mejor resultado con RNAs** se consiguió con 11 características y utilizando el **método de entrenamiento de descenso por gradiente** con función de coste error cuadrático medio. Se consiguió alcanzar un **porcentaje de acierto del** 96,9 %. Los mismos experimentos realizados con el discriminante lineal de Fisher y el algoritmo k-NN no mejoraron el resultado anterior.

Los **resultados** obtenidos son **altamente satisfactorios** puesto que **mejoran algunos de los obtenidos en otros trabajos** para la clasificación voz/no-voz [5, 6]. Además, cabe reseñar que utilizando como **idea innovadora**, un **clasificador en etapas** para la clasificación de los entornos de conversación en ambiente tranquilo (voz limpia), conversación en ambiente ruidoso (voz con ruido) y ruido y música (no voz), se consiguen **mejores resultados** los que se obtendrían con un sólo clasificador que discriminara directamente entre las tres clases [7].

Como conclusión final, se puede afirmar que se han alcanzado todos los objetivos que se perseguían con la realización del presente proyecto, logrando numerosos progresos adicionales y abriendo la puerta a nuevos avances en esta línea de investigación.

2. Originalidad

La principal innovación de este proyecto consiste en utilizar un **clasificador, en etapas, y de baja carga computacional para la detección automática de entornos sonoros en audífonos digitales**. En este proyecto, se propone un sistema de clasificación en etapas para la discriminación de los tres principales entornos sonoros considerados: voz en ambiente tranquilo, voz en ambiente ruidoso y voz y música. Debido a que en los audífonos digitales, la importancia de cada clasificación es distinta, el sistema propuesto está basado en una estrategia de *divide y vencerás*, con objeto de obtener un sistema más robusto. Para ello, en la primera etapa de clasificación, se clasifica entre voz/no-voz, conteniendo esta última clase ruido y música; y en otra segunda etapa, si la señal se clasificó como voz, se trata de discriminar entre voz limpia/voz con ruido. El objetivo final es realzar la voz en entornos acústicos cambiantes, minimizando el ruido de fondo. La originalidad estriba en diminuir el coste computacional en el dispositivo manteniendo (e incluso aumentando) la probabilidad de clasificar adecuadamente el entorno sonoro.

Existen muy pocos trabajos precedentes de investigación en ingeniería de clasificación voz/no-voz [8,9]. Además, recientemente se han realizado diversos esfuerzos reseñables en la búsqueda de las mejoras características para esta tarea [5,6,10]; sin embargo, son muy pocas las referencias que se encuentran de cara a una clasificación de entornos sonoros aplicada a la detección automática de entornos en audífonos digitales basados en DSPs.

Uno de los trabajos más significativos que encontramos en la literatura para la discriminación de entornos sonoros aplicada a audífonos digitales [11] propone un sistema de reducción de ruido basado en un proceso de identificación voz/no-voz utilizando cuatro características. Aunque los resultados de los tests de inteligibilidad realizados a personas que sufren pérdidas de audición mostraban una buena reducción de ruido, el sistema no es capaz de detectar forma automática el entorno en el que se encuentra el paciente para activar el programa que aumente la calidad subjetiva del audio que escucha el paciente.

La originalidad del proyecto ha sido validada en diversos foros científicos especializados, por medio de **publicaciones en revistas internacionales con índice de impacto y en congresos nacionales e internacionales**, tal y como se detalla en el Anexo II.

3. Resultados

En este apartado se presentan los resultados más importantes obtenidos en el proyecto. Para una mayor claridad, los resultados de cada una de las tareas de clasificación se mostrarán por separado.

1. **Clasificación voz/no-voz**

   El objetivo de esta primera tarea es clasificar la señal de audio de entrada como voz o no-voz. Las señales de voz contienen tanto conversaciones en ambiento tranquilo (voz limpia) como conversaciones en ambiente ruidoso (voz con ruido), mientras que las señales de no-voz incluyen música (vocal e instrumental) y ruido. A continuación se describen los experimentos ligados a los resultados.

   Los primeros experimentos llevados a cabo en la tarea de clasificación voz/no-voz fueron realizados con una **base de datos** reducida que constaba de 546 **señales de audio**. Se aplicaron dos métodos de clasificación simples como son el discriminante lineal de Fisher y el algoritmo k-NN. Como primera aproximación y para aislar el efecto de cada parámetro del resto lo más posible, se realizaron **experimentos** de clasificación **utilizando cada uno de los parámetros de forma aislada**. Para el algoritmo **k-NN (con k igual a 4)**, el **mejor resultado** se obtuvo con los **MFCC obteniendo un porcentaje de acierto ligeramente inferior al 90 %**. Sin embargo, para el discriminante lineal de Fisher, la característica SFM es la que presenta una mayor tasa de acierto, en torno al 84 %. Si se **combinan las características** que de forma aislada proporcionan una mejor separabilidad, se ha podido comprobar que es posible **incrementar la tasa de acierto**, **obteniendo un 94,5 %** en el caso de utilizar un clasificador basado en el **algoritmo k-NN (con $k = 4$)** y las características Spectral Centroide, Voice2White, ZCR, LSTER, SFM, Loudness y los 10 primeros coeficientes MFCC. Por último, si se utilizan todas las características de forma conjunta, el porcentaje de acierto para un clasificador basado en el discriminante lineal de Fisher se incrementa hasta un 93.0 %, mientras que para un clasificador basado en el algoritmo k-NN, no aumenta con respecto al obtenido con el conjunto anterior.
   Con objeto de mejorar los resultados obtenidos hasta el momento, se decidió aplicar la técnica de **combinación de clasificadores**, realizando distintos experimentos con distintos conjuntos de características tanto para clasificadores basados en el discriminante lineal de Fisher como para clasificadores basados en el algoritmo k-NN. Al final, el mejor resultado se obtuvo con el sistema propuesto en la Fig. 3 alcanzado una **tasa de acierto del 97 %**.

   En base a obtener resultados más generalizados, se evaluó la eficiencia de las RNAs para la discriminación voz/no-voz. Para estos experimentos se utilizó la base de datos diseñada para el proyecto y que constaba de 2936 señales de audio. Para la realización de los experimentos se consideraron tres algoritmos de entrenamiento distintos como son: el método de descenso por gradiente con dos funciones objetivo distintas (error cuadrático medio y entropía), el método Levenberg-Marquardt y el método Levenberg-Marquardt con técnicas de regularización bayesiana. Para cada uno de los métodos, se realizaron experimentos con cada característica por aislado, así como para distintos conjuntos de características. Para cada experimento se realizó un barrido de neuronas en la capa oculta de 1 a 30, repitiendo cada experimento 10 veces en aras de obtener un resultado promedio. Las tablas 1 y 2 muestran el mejor resultado obtenido para cada método de entrenamiento con la mejor característica y el conjunto total de características, respectivamente. Asimismo, en la tabla 2 se muestra el número de neuronas necesario en la capa oculta.

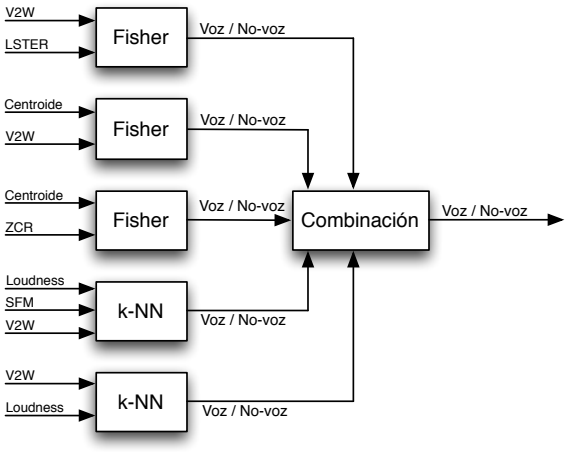

Figura 3: Esquema del sistema de combinación de clasificadores con las características utilizadas en cada clasificador. Se utilizó una forma sencilla de combinación de clasificadores como es la del *voto por mayoría*.

Método de entrenamiento	**Función objetivo**	**Característica**	**Error de validación**	**Tasa acierto (%)**
Descenso grad.	mse	Centroide	0.0728	86.6
	entropía	MFCC(20 coeficientes)	0.2303	**88.4**
Levenberg-Marquardt		SFM	0.0654	88.2
Regularización Bayesiana		SFM	28.5704	87.3

Tabla 1: Error de validación y tasa de acierto (%) para la mejor característica de cada uno de los métodos de entrenamiento utilizados en clasificación voz/no-voz.

Como se puede observar, el mejor resultado se obtiene con el **método Levenberg-Marquardt con técnicas de regularización bayesiana** cuando se utilizaban todas las características, logrando un **porcentaje de acierto del** 96,6 %. Los mismos experimentos se repitieron para el discriminante lineal de Fisher y el algoritmo k-NN, obteniendo siempre peores resultados.
Asimismo, se aplicó la técnica VSL (Variable Structure Learning) encontrada en la literatura para hallar el número óptimo de neuronas en la capa oculta de la RNA. Aunque en el proyecto se muestran los resultados obtenidos al aplicar esta técnica a distintos conjuntos de características con el método descenso por gradiente con función objetivo error cuadrático medio, aquí y únicamente a modo indicativo se muestra en la Fig.4 como para 2 características a la entrada, a partir de 4 neuronas en la capa oculta, el error no disminuye más. Esto significa que el tamaño óptimo de la red para 2 características a la entrada es de 4 neuronas en la capa oculta.

2. **Voz limpia/Voz con ruido**

   El objetivo de esta segunda tarea es distinguir entre voz en ambiente tranquilo (voz limpia) y voz en ambiente ruidoso (voz con ruido). Para realizar los experimentos se utilizaron RNAs y la base de datos de 2936 señales de audio. Se consideraron los mismos algoritmos

Método de entrenamiento	Función objetivo	Número de neuronas	Error de validación	Tasa acierto (%)
Descenso grad.	mse	24	0.0363	96.0
	entropía	20	0.0563	95.8
Levenberg-Marquardt		13	0.0364	93.7
Regularización Bayesiana		24	14.2077	**96.6**

Tabla 2: Error de validación y tasa de acierto (%) para todas las características con cada uno de los métodos de entrenamiento utilizados en clasificación voz/no-voz.

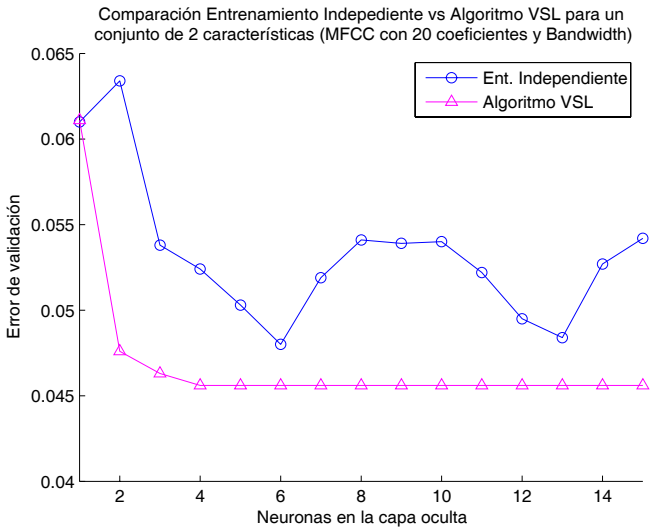

Figura 4: Comparación del error de validación entre aplicar o no el algoritmo VSL para un conjunto de 2 características (MFFC con 20 coeficientes y Bandwidth) y un número de neuronas en la capa oculta entre 1 y 15.

de entrenamiento que en el caso de la clasificación voz/no-voz. La mayor experimentación se realizó para el método de descenso por gradiente y función objetivo error cuadrático medio con el que se calculó el error de validación y tasa de acierto de cada una de las características utilizadas de forma individual, y se fueron probando distintas combinaciones de características, seleccionando aquellas que de forma aislada presentaban menor error de validación. A modo de resumen, en la tabla 3 se muestran los resultados más significativos.

Destacar que el **mejor porcentaje de acierto**, **96.9 %, se obtiene para el método descenso por gradiente** y con función objetivo el **error cuadrático medio con** 11 **características.**

Asimismo, se aplicó la técnica VSL a los conjuntos de la tabla 3, utilizando como método de entrenamiento el descenso por gradiente y función objetivo el error cuadrático medio. Se puedo comprobar que era posible reducir el tamaño de la red, manteniendo un nivel aceptable de aciertos.
Finalmente y para contrastar los resultados obtenidos, se realizaron experimentos similares con el discrimiante lineal de Fisher y el algoritmo k-NN. Con ninguno de los dos métodos se cosiguió una porcentaje de acierto superior al 96,9 %.

	Método de entrenamiento		Número de neuronas	Error de validación	Tasa acierto (%)
5 caract.	Grad.	mse	5	0.0553	94.7
		entropía	13	0.1672	**95.7**
	Regularización		6	8.9864	95.0
	Levenberg-Marquardt		3	0.0462	91.5
	Método de entrenamiento		**Número de neuronas**	**Error de validación**	**Tasa acierto (%)**
11 caract.	Grad.	mse	9	0.0510	**96.9**
		entropía	16	0.1535	96.0
	Regularización		3	8.4582	95.9
	Levenberg-Marquardt		5	0.0429	92.5
	Método de entrenamiento		**Número de neuronas**	**Error de validación**	**Tasa acierto (%)**
14 caract.	Grad.	mse	3	0.0448	95.7
		entropía	20	0.1407	**96.4**
	Regularización		14	7.6366	95.9
	Levenberg-Marquardt		10	0.0411	94.0

Tabla 3: Comparación de resultados en clasificación voz limpia/voz con ruido entre los distintos algoritmos de entrenamiento para los conjuntos de 5, 11 y 14 características.

4. Aplicabilidad

La aplicabilidad de este proyecto tiene dos vertientes claramente diferenciadas: una centrada en la investigación y otra, más comercial, orientada al desarrollo de prótesis auditivas que realcen de forma selectiva la voz, minimizando los efectos de ruido de fondo. **Desde el punto de vista de la investigación, este proyecto constituye el punto de partida sobre el que se desarrollará la futura tesis de la proyectista**. Tesis cuyos objetivos están íntimamente ligados con este proyecto

1. Estudios de métodos de selección de características, así como definición de nuevas características *ad-hoc* que permitan aumentar la separabilidad entre entornos.
2. Estudio de distintos algoritmos para la clasificación automática de señales sonoras.
3. Análisis del efecto de la precisión finita del DSP sobre el que se implementará el sistema.
4. Implementación en tiempo real sobre un DSP de propósito específico, a fin de obtener un primer prototipo totalmente operativo del audífono.

Desde el punto de vista comercial, supone una innovación en la detección y selección automática de entornos en audífonos digitales, al utilizar un clasificador en etapas y estudiar la posibilidad de implementar una RNA como clasificador. De hecho, el grupo de investigación dentro del cual la proyectista ha realizado este trabajo, realizó un prototipo comercial de audífono digital para la empresa ARIX TELECOM. Con este proyecto, y el consecuente trabajo a realizar en la tesis de la proyectista, se pretende desarrollar un nuevo prototipo de audífono digital que permita detectar el entorno sonoro en el que se encuentra el paciente y seleccionar el programa adecuado para la situación acústica detectada.

Por tanto, se trata de un proyecto práctico e innovador cuyos resultados tienen aplicación tanto comercial como de investigación. La primera de ellas de cara a desarrollar un primer prototipo funcional de un audífono digital que incorpore técnicas

avanzadas de procesado de señal y que sea capaz de adaptarse de forma automática al entorno sonoro en el que se encuentre el paciente, y la segunda como herramienta de investigación realizada en el ámbito de una tesis doctoral.

Desde el punto de vista social, el proyecto contribuye al desarrollo de audífonos que puedan ser utilizados por personas que, como los ancianos, tienen dificultades para manejar correctamente los audífonos presentes actualmente en el mercado. La clasificación automática objeto de este proyecto les permite mejorar sus habilidades de comunicación y aumentar su grado de bienestar.

Referencias

[1] M. Büchler, "Algorithms for sound classification in hearing instruments," Ph.D. dissertation, Swiss Federal Institute of Technology, Zurich, 2002.

[2] G. Keidser, "The relationships between listening conditions and alterative amplification schemes for multiple memory hearing aids," *Ear Hear*, vol. 16, pp. 575–586, 1995.

[3] ——, "Selecting different amplification for different listening conditions," *J. of the American Academy of Audiology*, vol. 7, pp. 92–104, 1996.

[4] C. Bishop, *Neural Networks for Pattern Recognition*. Clarendon Press. Oxford, 1995.

[5] H. Meinedo and J.Ñeto, "A stream-audio segmentation, classification and clustering pre-processing system for broadcast news using ann models." in *InterSpeech 2005*, 2005.

[6] R. Huang and J. H. Hansen, "High-level feature weighted gmm network for audio stream classification," 2004.

[7] E. Alexandre, L. Alvarez, L. Cuadra, M. Rosa, and F. Lopez, "Automatic sound classification algorithm for hearing aids using a hierarchical taxonomy," in *New Trends in Audio and Video*, A. Dobrucki, A. Petrovsky, and W. Skarbek, Eds. Politechnika Bialostocka, 2006, vol. 1.

[8] K. T. H. Kobetake and A. Ishida, "Speech/non-speech discrimination for speech recoginition system under real life noise environments," in *IEEE ICASSP*, 1989, pp. 365–386.

[9] M. Pwint and F. Sattar, "A new speech/non-speech classification method using minimal walsh basis functions," in *IEEE International Symposium on Circuits and Systems (ISCAS 2005)*, vol. 3, 2005, pp. 2863–2866.

[10] Y.-K. L. Won-Ho Shin, Byoung-Soo Lee and J.-S. Lee, "Speech/non-speech classification using multiple features for robust endpoint detection," *ICASSP*, pp. 1399–1402, 2000.

[11] K. Itoh and M. Mizushina, "Environmental noise reduction based on speech/non-speech identification for hearing aids," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1997)*, vol. 3, 1997, pp. 419–422.

Anexo I. Financiación

Este proyecto ha sido realizado en el marco del proyecto de investigación "Mejora de la inteligibilidad de la voz en audífonos digitales" (CAM-UAH-2005/036), cofinanciado por la Comunidad de Madrid y la Universidad de Alcalá.

Anexo II. Publicaciones

Como muestra de la innovación y originalidad de los resultados obtenidos, a continuación se muestra una relación de publicaciones derivadas de este proyecto.

Revistas internacionales

1. E. Alexandre, L. Álvarez, M. Rosa and F. López, "Two-layer automatic sound classification system for conversation enhacement in hearing aids". Integrated Computer Aided Engineering. Vo. 15, No. 1, 2008.

2. E. Alexandre, L. Cuadra, L. Álvarez, M. Rosa and F. López, "Automatic sound classification for improving speech intelilligibility in hearing aids using a layered structure". Lecture Notes in Computer Science. 2006.

3. E. Alexandre, L. Cuadra, L. Álvarez, M. Rosa, F. López. "Noise-assisted spectral centroid for sound classification in hearing aids". IEEE Signal Processing Letters. Enviado y sujeto a revisión.

Capítulos de libro

1. E. Alexandre, L. Álvarez, L. Cuadra, M. Rosa and F. López. "Automatic sound classification algorithm for hearing aids using a hierarchical taxonomy". New Trends in Audio and Video, Vol. 1, 2006.

Congresos internacionales

1. L. Álvarez, E. Alexandre, R. Vicen, L. Cuadra and M. Rosa, "A Constructive Algorithm for Multilayer Perceptrons for Speech/Non-Speech Classification in Hearing Aids" aceppted for AES 124th Convention, Amsterdam, 2008.

2. E. Huerta, E. Alexandre, R. Gil, L. Álvarez, J. Amor. "Analysis of the effects of finite precision in sound classifiers for digital hearing aids". Accepted for AES 124th Convention. Amsterdam, 2008.

3. J. Amor, E. Alexandre, R. Gil, L. Álvarez, E. Huerta. "Music-inspired harmony-search algorithm applied to feature selection for sound classification in hearing aids". Accepted for AES 124th Convention, Amsterdam, 2008.

4. L. Álvarez, E. Alexandre, L. Cuadra and M. Rosa, "On the training of Multilayer Perceptrons for Speech/Non-Speech Classification in Hearing Aids". AES 122nd Convention, Preprint 7136 , 5-8 May 2007, Vienna.

5. E. Alexandre, L. Álvarez, L. Cuadra, M. Rosa and F. López. "Automatic sound classification algorithm for hearing aids using a hierarchical taxonomy". Conferencia invitada en el XI Symposium AES New Trends in Audio and Video, 2006, Bialystok.

6. E. Alexandre, L. Álvarez, L. Cuadra, M. Utrilla. "Exploring the feasibility of a two-layer NN-based sound classifier for hearing aids". EUSIPCO, 2007.

7. E. Alexandre, L. Cuadra, L. Álvarez, M. Rosa. "NN-based automatic sound classifier for digital hearing aids". IEEE WISP, 2007.

8. E. Alexandre, R. Gil, L. Cuadra, L. Álvarez, M. Rosa. "Speech/music/noise classification in hearing aids using a two-layer classification system with MSE linear discriminants". EUSIPCO 2008.

9. L. Cuadra, E. Alexandre, L. Álvarez, M. Rosa. "Reducing the computational cost for sound classification in hearing aids by selecting features via genetic algorithms with restricted search". IEEE International Conference on Audio, Language and Image Processing. Shanghai, 2008.

Congresos nacionales

1. L. Álvarez Pérez, E. Alexandre Cortizo, L. Cuadra Rodríguez, R. Gil Pita y F. López Ferreras, "Discriminación automática voz/no-voz mediante combinación de clasificadores". XXI Simposium Nacional de la Unión Científica Internacional de Radio (URSI 2006), pp. 1729-1735 , 12-15 Septiembre 2006, Oviedo.