Diseño de un sistema de reconocimiento del habla mediante electromiografía

E. López Larraz[1], O. Martínez Mozos[1], J.M. Antelis Ortiz[1], J. Damborenea Tajada[2], J. Mínguez Zafra[1]

[1] Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, Zaragoza, España,
{525242, ommozos, antelis, jminguez}@unizar.es

[2] Servicio de Otorrinolaringología, Hospital Miguel Servet, Zaragoza, España

Resumen

Este trabajo presenta un sistema de reconocimiento del habla en idioma español usando señales electromiográficas de los músculos de la cara. Nuestro sistema analiza las señales de 8 músculos en paralelo mientras el paciente pronuncia los fonemas correspondientes a 30 silabas del idioma español. Para cada grupo de señales se calculan un conjunto de características que representan el sonido silábico. Estas características se utilizan para entrenar un clasificador multiclase para las 30 sílabas. El clasificador es capaz de reconocer una nueva pronunciación con un porcentaje de efectividad de, prácticamente, un 71%, superando los ratios para la clasificación de fonemas en otros idiomas presentados en trabajos anteriores.

1. Introducción

Uno de los tipos de interfaz entre persona y ordenador en el que más se ha investigado en los últimos años son los sistemas de reconocimiento del habla (en inglés automatic speech recognition, o ASR), que se basan en grabar las señales de voz producidas por una persona, procesarlas y clasificarlas para reconocer qué es lo que se ha dicho. Sin embargo estos interfaces presentan dos graves inconvenientes que pueden hacer inservible un sistema de este tipo:

- En un entorno ruidoso, un ASR no resultará útil, ya que mezclará las señales de voz con el ruido y no se conseguirá que las reconozca adecuadamente.
- Para personas con cualquier tipo de discapacidad en el habla, estos sistemas tampoco resultan valiosos, dado que no serán capaces de identificar las señales de voz.

Para salvar estos inconvenientes se han empezado a desarrollar interfaces de reconocimiento del habla basados en electromiografía (EMG). Con esta tecnología no es necesario que el usuario produzca señales acústicas, ya que el simple movimiento de los músculos faciales produce un voltaje que puede ser detectado y medido con pequeños electrodos para después procesarlo en un computador.

En los últimos años se han estudiado sistemas automáticos de reconocimiento de fonemas y palabras en otros idiomas usando señales EMG ([1], [2], [3]). Sin embargo no tenemos constancia de ningún sistema de reconocimiento de sonidos del idioma español usando EMG. Por ello pensamos que el sistema presentado en este trabajo es uno de los primeros en reconocer un subconjunto de sílabas del idioma español. Por otra parte, el porcentaje de reconocimiento que obtenemos en nuestro sistema con un 71% de efectividad en la clasificación supera los ratios de clasificación en fonemas presentados para otros idiomas.

2. Estudio fisiológico

2.1. Estudio de la musculatura facial

El estudio de los músculos que componen la anatomía facial es un punto clave para la obtención de unas señales adecuadas que permitan una correcta clasificación posterior. Sobre su superficie se colocarán los electrodos para la adquisición de los datos, por tanto, una buena elección de los músculos, junto con una correcta colocación de los sensores proporcionará unas muestras suficientemente buenas con las que poder trabajar.

Por ello, tras varias pruebas se ha decidido emplear una estrategia en la que se colocarán 16 electrodos que, configurados de manera bipolar, proporcionarán 8 canales. La tierra será otro electrodo colocado en el centro de la frente y todos ellos tendrán como referencia el lóbulo de la oreja. Los músculos sobre los que se situarán los electrodos serán: *Levator labii superioris*, *Zygomaticus major*, *Orbicularis oris*, *Risorius*, *Depressor anguli oris*, *Depressor labii inferioris*, vientre anterior del músculo *Digastric*, y por último la lengua. Estos músculos se han escogido porque todos ellos han sido utilizados en alguno de los trabajos referenciados y, además están distribuidos por toda la cara, no restringiendo así la colocación a una zona determinada de ésta. En la figura 1 podemos ver esta configuración.

2.2. Estudio del vocabulario

Dado un claro objetivo, como es el de diseñar una prótesis de reconocimiento del habla, necesitamos estudiar la naturaleza de ésta para comprender cómo se forma y qué manera será la idónea para desarrollar el mecanismo reconocedor. Pese a que el propósito final es el de abarcar todo el lenguaje castellano, el primer paso antes de reconocer sus palabras es tratar de reconocer las sílabas que las componen. Así, siguiendo este orden lógico, primero se reconocerán estas unidades básicas del habla, posteriormente serán agrupadas en palabras y, más adelante, se buscará obtener un reconocimiento semántico que dote de sentido a las frases formadas.

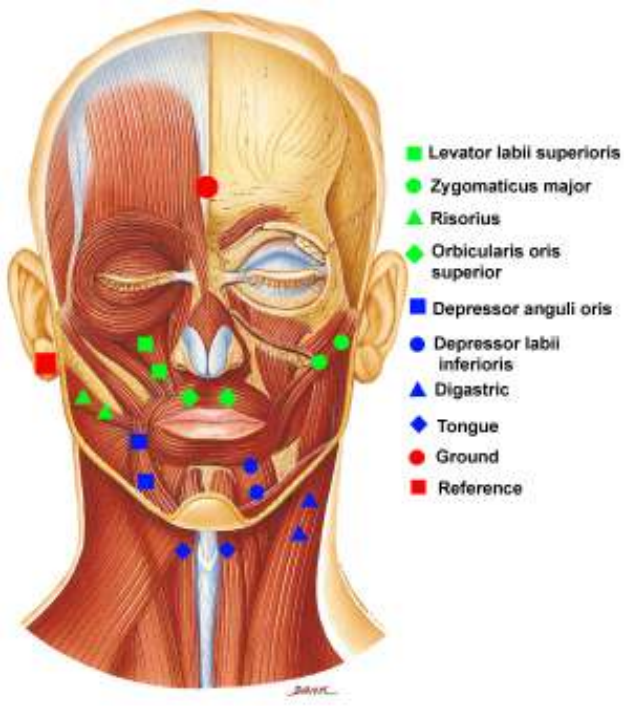

Figura 1. *Mapa de la musculatura facial, en el que se han indicado con distintas formas y colores dónde se colocarán los sensores*

Una vez en este punto debemos definir cuáles son las sílabas que queremos incluir en nuestro sistema. Deberán ser un subconjunto de las sílabas más simples, compuestas por una consonante seguida de una vocal y buscaremos las que, de alguna manera, sean suficientemente representativas. Según la forma en la que se articulan, este tipo de sílabas están divididas en 5 grupos principales: *labiales*, *dentales*, *palatales*, *velares* y *alveolares* [4].

Para constituir el vocabulario que deberá reconocerse posteriormente, se han tomado las 5 vocales y la combinación de éstas con una consonante perteneciente a cada grupo mencionado. La tabla 1 muestra cuáles son las 30 sílabas seleccionadas.

Vocales	*A*	*E*	*I*	*O*	*U*
Labiales	*PA*	*PE*	*PI*	*PO*	*PU*
Dentales	*TA*	*TE*	*TI*	*TO*	*TU*
Palatales	*YA*	*YE*	*YI*	*YO*	*YU*
Velares	*KA*	*KE*	*KI*	*KO*	*KU*
Alveolares	*LA*	*LE*	*LI*	*LO*	*LU*

Tabla 1. *Vocabulario diseñado para su posterior reconocimiento*

3. Métodos

El prototipo debe estar compuesto por dos bloques principales. El primero de ellos es el sistema de tratamiento de señales, que transformará los datos EMG, aplicando operaciones o transformaciones, en una serie de valores. El segundo de los sistemas será una máquina de aprendizaje que tomará como entrada esos valores y, mediante algoritmos de clasificación, tratará de reconocer a qué sílaba corresponde cada patrón.

3.1. Tratamiento de señales

La extracción de características se ha realizado mediante scripts desarrollados en Matlab. Para cada uno de los 8 canales, los valores que se calculan son los siguientes:

- Fast Fourier Transform (FFT)
- Root Mean Square (RMS)
- Amplitud media de la señal
- Amplitud máxima
- Kurtosis
- Mel Frequency Cepstral Coefficients (MFCC)
- Integrated Absolute Value (IAV)
- Zero Crossing
- Suma de la señal
- Suma de la señal rectificada

De la FFT se toman las 20 primeras frecuencias y del MFCC se utilizan 13 coeficientes, así que en total se tienen 41 valores para cada canal. Estos se concatenan a la hora de clasificar, utilizándose por tanto 328 valores para codificar cada muestra.

3.2. Sistema de clasificación

La máquina de aprendizaje empleada para realizar las distintas clasificaciones ha sido AdaBoost combinado con un árbol de decisión. Con esto se han propuesto cuatro esquemas para tratar de distinguir, de la mejor manera posible, las 30 sílabas.

El primer esquema que es el más sencillo e intuitivo, es un clasificador multiclase estándar, y consiste en entrenar la máquina de aprendizaje con ejemplos de todas las clases existentes, en este caso 30.

El siguiente es un esquema matricial, que consiste en una división simple del problema en dos más pequeños. Se utilizará un clasificador para tratar de identificar el comienzo de las sílabas (clasificador por filas) y otro para reconocer las terminaciones (clasificador por columnas). Así, el primero de ellos tendrá 6 clases (Vocal, P~, T~, Y~, K~ y L~) y el segundo, 5 (~A, ~E, ~I, ~O y ~U).

Por último se han propuesto dos esquemas cuyo funcionamiento es similar al matricial. El primero de ellos utiliza el clasificador de comienzos, igual que en el caso anterior, pero, después emplea 6 clasificadores distintos específicos según el resultado que genere el primero. Así si el clasificador por filas determina que una muestra comienza por *P*, seguidamente se utiliza un clasificador que únicamente distinga entre las sílabas *PA, PE, PI, PO, PU*. El otro esquema trabaja de manera análoga, pero ejecuta primero el clasificador por columnas y después utiliza uno de los 5 clasificadores especializados en

distinguir entre los posibles comienzos, según el resultado que haya dado el primero.

3.3. Experimentación

Las señales electromiográficas se han obtenido en una sesión de experimentación en la que se adquirieron 150 muestras de cada sílaba.

Los electrodos que se han utilizado son de la marca Grass, con un diámetro de 10mm y fabricados en oro. Se colocaron con una configuración bipolar como la mostrada en la figura 1 y se conectaron a un amplificador que, mediante un cable USB 2.0, se comunica con el computador. La frecuencia de muestreo para registrar los datos ha sido de 2.400 Hz, las señales adquiridas han sido filtradas paso-banda mediante un filtro implementado en el hardware del amplificador, de 5 a 500 Hz, ya que es en esa banda de frecuencias donde se halla la información más importante. También se ha empleado un *notch-filter*, para evitar el rango de frecuencias de 48 a 52 Hz, debido a que en esa banda se produce ruido por la interacción con la instalación eléctrica del edificio.

Para la interfaz persona-ordenador se ha empleado una plataforma de código abierto llamada BCI2000 [5], que permitió la sincronización del amplificador con una serie de estímulos visuales, lo que facilitó en gran medida la adquisición controlada de los datos EMG.

4. Resultados

A continuación se mostrarán los resultados obtenidos por los cuatro esquemas de clasificación que se explicaron en el apartado 3.2. Para realizar todas las pruebas se ha empleado un software llamado Weka [6], que proporciona implementaciones de varios algoritmos de clasificación, que pueden ser fácilmente aplicados a cualquier conjunto de datos.

Como ya se ha mencionado, los esquemas de clasificación se han diseñaron utilizando como clasificador AdaBoost [7] combinado con el árbol de decisión J4.8, que es la implementación que utiliza Weka del árbol C4.5 [8].

Los porcentajes que se mostrarán han sido calculados utilizando *10 fold cross-validation*, lo que garantiza unos resultados más fiables por las 10 validaciones que realiza, utilizando cada muestra 9 veces para entrenar y una para clasificar.

4.1. Clasificador de 30 clases

Para el clasificador multiclase estándar, el resultado medio conseguido es del **70,93%.** La matriz de confusión de la figura 2 muestra el porcentaje de aciertos obtenidos para cada sílaba. Está estructurada ordenando las sílabas por terminación (*A, PA, TA, YA, KA, LA, E, PE, TE, YE, KE, LE…*). Como puede observarse, la mayoría de las confusiones se dan entre sílabas con igual terminación (*TA, YA, KA, LA; TE, YE, KE, LE…*), sin embargo, las sílabas vocálicas y las que empiezan por *P* se distinguen con mayor facilidad que el resto, algo que, seguramente, está causado porque son los dos únicos grupos en los que no se emplea la lengua para su gesticulación. Las líneas paralelas a la diagonal principal que aparecen son debidas a las confusiones que se producen por terminaciones similares; las más significativas son las que aparecen entre las terminaciones *E-I*, aunque también surgen, en menor medida entre *O-U* y alguna entre *A-E*.

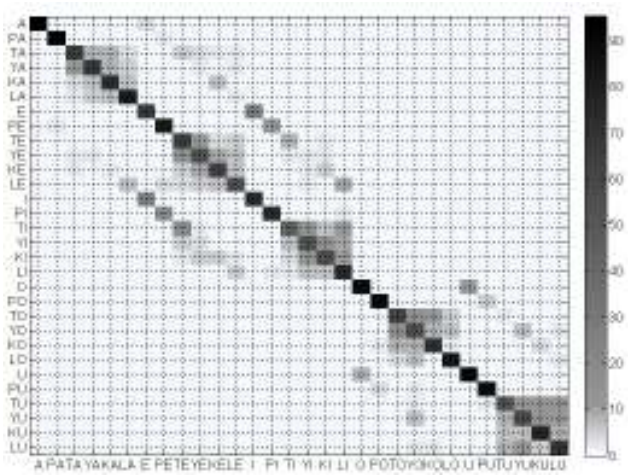

Figura 2. *Matriz de confusión del clasificador correspondiente al clasificador de 30 clases. Colores más oscuros indican un mayor índice de reconocimiento.*

4.2. Clasificador matricial

La probabilidad de acierto que proporciona el esquema matricial se obtiene multiplicando las probabilidades medias de cada uno de los clasificadores que lo componen. La media del clasificador por comienzos es del 77,47%, mientras que la del que actúa por terminaciones es del 87,33; así la media total conseguida es del **67,65%**. La figura 3 muestra la matriz de confusión del clasificador por comienzos, como puede verse se producen las mismas confusiones que se veían en el caso anterior, agrupándose por un lado las sílabas que empiezan por *T, Y, K, L* y distinguiéndose con más claridad las vocales y las que empiezan por *P*.

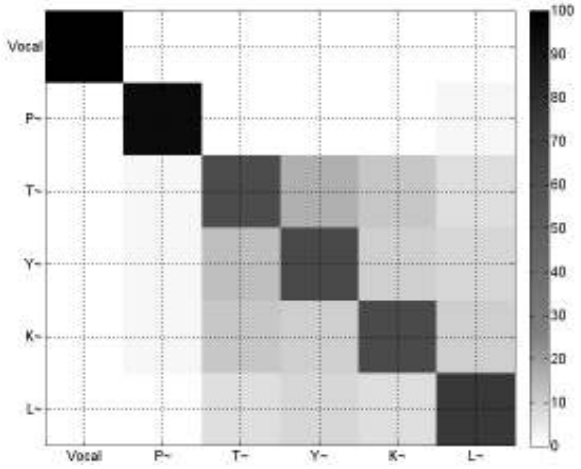

Figura 3. *Matriz de confusión esquema matricial. Colores más oscuros indican un mayor índice de reconocimiento.*

4.3. Clasificadores condicionales

Para los dos esquemas condicionales el resultado obtenido es muy similar. Para realizar el cálculo de la media de

acierto del esquema que usa un clasificador de comienzos estándar y 6 clasificadores condicionales de terminación, se calcula primero el promedio de los 6 clasificadores condicionales y, después, se multiplica por el porcentaje medio del clasificador de comienzos. Para el esquema que usa el clasificador de terminaciones estándar y los 5 clasificadores de comienzo condicionales se calcula del mismo modo: primero la media de los condicionales y después se multiplica por el clasificador de terminaciones. Los resultados obtenidos son de un **68,89%** de aciertos en el primer esquema y un **69,48%** el segundo.

La figura 4 corresponde a la matriz de confusión del clasificador condicional que actúa después de que el clasificador de comienzos determine que se trata de una sílaba que empieza por *P,* distinguiendo solamente las sílabas *PA, PE, PI, PO, PU*. Puede verse que la mayoría de las confusiones se producen entre las terminaciones *E-I,* y en menor medida las *O-U.*

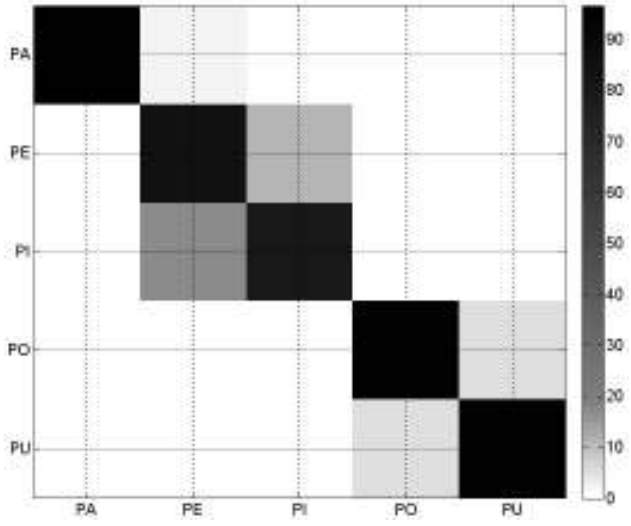

Figura 4. *Matriz de confusión del clasificador condicional de terminaciones que actúa cuando el comienzo de una sílaba es una P. Colores más oscuros indican un mayor índice de reconocimiento.*

5. Conclusiones y trabajo futuro

Los resultados obtenidos son muy buenos si se tiene en cuenta que un clasificador aleatorio proporcionaría un resultado medio de un 3.3% de acierto. Además la mayoría de los errores que se producen en los clasificadores son razonablemente aceptables, ya que las confusiones que más se repiten son entre sílabas que, si bien su pronunciación sonora es ya parecida, más todavía lo es su gesticulación (piénsese, por ejemplo, en las sílabas *TI* y *YI,* que son dos de las que más se confunden mutuamente).

No tenemos constancia de ningún otro sistema de reconocimiento del habla realizado en castellano, pero con respecto a los realizados en otro idioma, los resultados son bastante coherentes. El más similar de los estudiados es [3], en el que reconocían 18 fonemas ingleses con un resultado medio próximo al 50%.

Con respecto a la expansión del prototipo, deberá estudiarse al ampliar el vocabulario si los rendimientos se mantienen próximos en los 4 esquemas de clasificación, ya que, pese a que el primero es el que mejor resultado ofrece, también es el que más tiempo tarda en entrenar y clasificar. Por ello habría que sopesar si compensa su utilización o, por el contrario, es mejor escalar el problema a costa de una posible reducción de las prestaciones. En la figura 5 se muestra la comparativa de los resultados medios de clasificación ofrecidos por los cuatro esquemas.

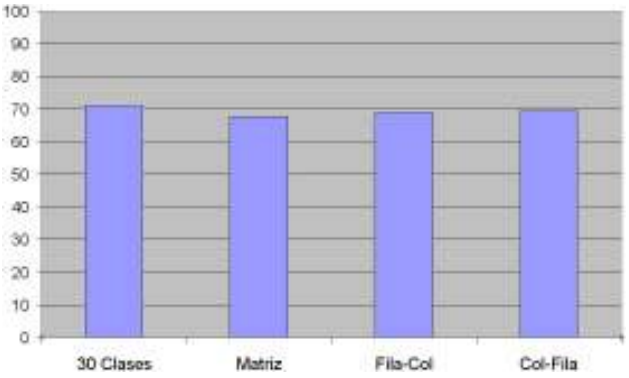

Figura 5. *Comparativa de los cuatro esquemas de clasificación*

Referencias

[1] S. P. Arjunan, H. Weghorn, D. K. Kumar, and Wai C. Yau. Vowel recognition of English and German language using Facial movement(SEMG) for Speech control based HCI. *HCSNet Workshop on the Use of Vision in HCI*, 2006.

[2] José AG Mendes, Ricardo R. Robson, Sofiane Labidi, and Allan Kardec Barros. Subvocal Speech Recognition Based on EMG signal Using Independent Component Analysis and Neural Network MLP. *Congress on Image and Signal Processing*, pages 221-224, 2008.

[3] Quan Zhou, Ning Jiang, Kevin Englehart, and Bernard Hudgins. Improved Phoneme-Based Myoelectric Speech Recognition. *IEEE Transactions on Biomedical Engineering,* 56(8), August 2009.

[4] Antonio Ríos Mestre. *La transcripción fonética automática del diccionario electrónico de formas simples flexivas del español: estudio fonológico en el léxico*. Laboratorio de Lingüística Informática de la Universidad Autónoma de Barcelona, 4 edition, 1999.

[5] G. Shalk, D.J. McFarland, T. Hinterberger, N. Birbaumer, and J.R. Wolpaw. BCI2000: A General-Purpose Brain-Computer Interface (BCI) System. *IEEE Transactions on Biomedical Engineering*, 51(6), May 2004.

[6] Ian H. Witten and Eibe Franck. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, 2 edition, 2005.

[7] Robert D. Vincent, Joelle Pineau, Philip de Guzman, and Massimo Avoli. Recurrent Boosting for Classification of Natural and Synthetic Time-Series Data. *Lecture Notes on Artificial Intelligence*, pages 192-203, 2007.

[8] J. R. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers, 1993.