Clasificación de patologías mediante análisis acústico de la voz

Y. Sabuco García [1], D. Ruiz Fernández [1,2]

[1] Escuela Politécnica Superior, Universidad de Alicante, Alicante, España, ysg4@alu.ua.es
[2] Escuela Politécnica Superior, Universidad de Alicante, Alicante, España, druiz@gcloud.ua.es

Resumen

Son múltiples los factores que intervienen en el acto de la fonación, por tanto, la voz puede evidenciar la presencia de enfermedades de diversa índole. El desarrollo de nuevos sistemas diagnósticos basados en el análisis de la voz puede ofrecer ventajas como la sencillez y la inocuidad. En este trabajo, se presenta un método de detección de sujetos patológicos a partir de la señal acústica de la voz mediante el conjunto de características ComParE de openSMILE y el modelo de ensemble Bagging.

1. Introducción

La voz es energía acústica que ha sido convertida por las cuerdas vocales a partir de energía aerodinámica que genera el aparato respiratorio [1]. Dada la gran cantidad de actores involucrados en la generación de la voz, directa o indirectamente, ésta puede ser el signo de enfermedades a diferentes niveles del organismo. Técnicas diagnósticas basadas en el análisis de voz podrían ofrecer ventajas como la inocuidad y la sencillez. Como consecuencia a esta cuestión, existen en la literatura múltiples estudios que analizan características mediante el análisis de la señal acústica a través modelos de inteligencia artificial con el fin de detectar e identificar patologías. En este sentido, se pueden encontrar referencias relacionadas con la detección de voces patológicas [2, 3] o la identificación de enfermedades como el cáncer de laringe [4], nódulos en laringe, edema de Reinke y parálisis de la laringe [2, 5], COVID-19 [6] o esclerosis lateral amiotrófica [7].

Uno de los factores limitantes en esta línea de investigación es la disponibilidad de las bases de datos y las diferencias que hay entre ellas. En ocasiones, las tareas vocales registradas o las frecuencias de muestreo son diferentes en función de la base de datos y la categorización de las patologías es muy dispar [8, 9]. Entre las bases de datos más utilizadas se encuentran la base de datos Massachusetts Eye and Ear Infirmary (MEEI) [10, 8, 2, 3, 4], la base de datos Saarbruecken Voice Database (SVD) [11, 8, 3, 12] y la base de datos Arabic Voice Pathology Database Samples (AVPD)[13, 14, 5].

En cuanto a los modelos elegidos para la detección e identificación de patologías por la voz, son numerosas las técnicas de inteligencia artificial utilizadas para la clasificación de enfermedades, se recurre tanto a modelos de aprendizaje profundo [12, 3, 4], como a otras técnicas de machine learning [7, 3, 2, 6].

Algo similar a la elección de los modelos ocurre con respecto al análisis de la señal acústica, son diversos los algoritmos y parámetros calculados en el dominio temporal, frecuencial y cepstral que se pueden extraer y, a día de hoy, no existe consenso con respecto a qué carácteristicas son las más propicias para el objetivo que se presenta. Entre los más utilizados en el dominio temporal se encuentran parámetros de frecuencia como la frecuencia fundamental y el jitter, parámetros de intensidad como el shimmer y parámetros de ruido como el Harmonic to Noise Ratio (HNR), Normaliced Noise Energy (NNE) o Noise to Harmonic Ratio (NHR) [7, 4]. Mientras que en el dominio cepstral, los coeficientes basados en la percepción del habla Mel Frequency Cepstral Coefficients (MFCC) son utilizados habitualmente [8, 12, 2].

Recientemente, han aparecido diversos estudios que han hecho uso de openSMILE [3, 6, 9]. Concretamente, el conjunto de características ComParE [15] es capaz de extraer un total de 6373 de modo automático. Son pocos los estudios que hayan utilizado este conjunto para la detección e identificación de patologías a través de la voz pese a las ventajas que puede suponer hacer uso de esta herramienta, que obtiene de una manera rápida y de un modo estandarizado tantas características. El único estudio encontrado que hace uso de este sistema para detección de voces patológicas, construye un conjunto de características con las tareas vocales /a/,/i/ y /u/ en diferentes todos y utiliza el modelo de aprendizaje supervisado Support Vector Machine (SVM)[16].

En este trabajo, el objetivo será detectar voces patológicas mediante la combinación de un método de extracción de características y un método de inteligencia artificial que no se ha utilizado antes para este propósito. Se utilizará la base de datos SVD, prestando especial atención en las limitaciones que se conocen acerca de ésta. La extracción de características será mediante el conjunto de datos ComParE de openSMILE y el método de ensemble Bagging.

2. Materiales y métodos

En el siguiente apartado se procederá a explicar la metodología llevada a cabo para preparar el conjunto de datos, extraer y pre-procesar las características, entrenar y validar el modelo de clasificación. En la siguiente figura se muestra un resumen de la metodología:

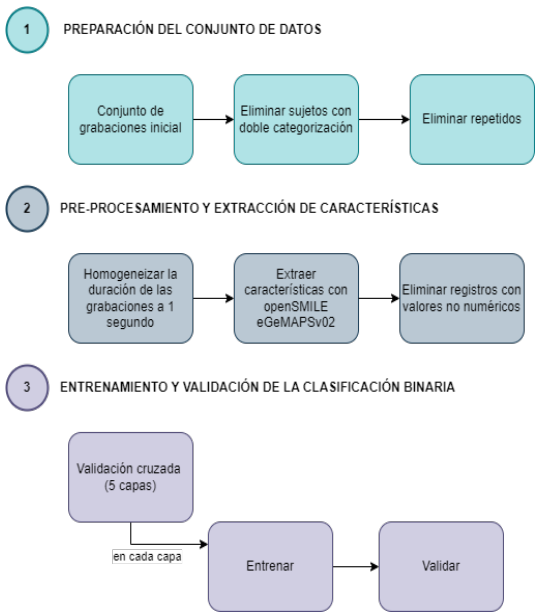

Figura 1. *Metodología*

2.1. Base de datos

Saarbruecken Voice Database es una base de datos abierta, distribuída por el Instituto de Fonética de la Universidad del Sarre, que posee en torno a 2000 sesiones donde se recopilaron grabaciones de las tareas vocales de los fonemas /a/, /i/ y /u/ producidas en tono normal, alto, bajo y con tono ascendente. Además, posee grabaciones de una frase en alemán y de la señal electroglotográfica (EGG). No obstante, y pese a la gran cantidad de registros que contiene, no todas las sesiones tienen todas las diferentes grabaciones recopiladas, contiene sujetos repetidos y existen individuos categorizados como sanos y patológicos al mismo tiempo[9]. Este factor no siempre es tenido en cuenta en los estudios que utilizan esta base de datos.

2.2. Preparación del conjunto de datos

Se utilizará la grabación de la tarea vocal /a/ en tono neutral de SVD cuya duración oscila de 1 a 3 segundos dependiendo del sujeto. Esta elección está basada en que es la única tarea que se recoge en las tres bases de datos más utilizadas en el estado del arte, y permitirá hacer comparativas mas precisas con trabajos anteriores. Estos registos poseen una frecuencia de muestreo de 50 KHz y una resolución de 16 bits.

Se parte con el número de 2041 grabaciones de este tipo (886 hombres y 1155 mujeres). Sin embargo, como se ha dicho en los apartados anteriores, puede ocurrir que los sujetos estén repetidos, y de estos sujetos repetidos, hay algunos categorizados al mismo tiempo como sanos y enfermos. Ante esta dificultad, se decide suprimir aquellos individuos con doble categorización, así como, eliminar los individuos repetidos. Además, en ocasiones los registros no disponen de la calidad suficiente y al realizar el análisis acústico se generan valores no numéricos que invalidan el conjunto de datos, estos registros también son suprimidos.

2.3. Pre-procesamiento y extracción de características

Para no obtener resultados dependientes de la duración se considerará sólo el primer segundo de cada grabación. Para la extracción de características se utilizará la herramienta open-source openSMILE, más concretamente, el conjunto de características denominado ComParE. Esta herramienta extrae automáticamente 6373 características. Su mecanismo se basa en extraer múltiples características de la señal acústica, lo que sus autores llaman descriptores de bajo nivel. A las cuales aplican descriptores funcionales, es decir, cálculos estadísticos como la media, la desviación estándar, cuartiles, etc.

Para la correcta validación de sistema se utilizará un método de validación cruzada de cinco capas. En este método, los datos se dividirán en cinco partes. Cada parte tendrá un conjunto de entrenamiento y uno de validación diferente que no se solapan de una iteración a otra. Habrá cinco iteraciones de entrenamiento independientes y el análisis del rendimiento se realizará con los datos de validación. La media de los datos de rendimiento de las cinco iteraciones corresponderán a los datos finales.

En la Figura 2, observamos el conjunto de datos resultante tras la preparación del conjunto de datos:

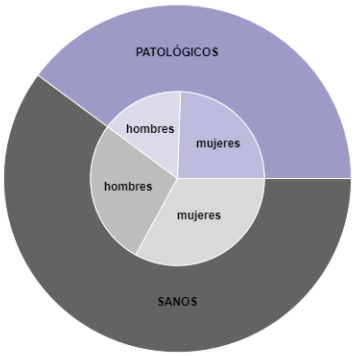

Figura 2. *Distribución del conjunto de datos final*

Conjunto de datos	Sanos	Patológicos	Total
Hombres	255	450	705
Mujeres	405	546	951
Total	660	996	1656

Tabla 1. *Distribución del conjunto de datos final*

2.4. Clasificación binaria

En un problema de clasificación, el objetivo del modelo es determinar a qué clase pertenece la entrada. Cuando se tienen dos clases objetivo se denomina, clasificación binaria. Para el problema de clasificación binaria propuesto en este estudio, discriminar entre sujetos sanos y patológicos, se utilizará un modelo de aprendizaje supervisado. Un modelo supervisado es aquel que resuelve

una tarea como la clasificación estableciendo una relación entre la entrada y la salida del modelo.

El método de aprendizaje supervisado utilizado en este estudio es el clasificador Bagging, este clasificador pertenece a los métodos ensemble o combinados. Estos métodos se basan en la combinación lineal de modelos que hacen una predicción simple, éstas se combinan para resolver un problema más complejo. En función de la combinación de los modelos simples que componen un método de ensemble podemos distiguir:

- Votación por mayoría: en éste método cada modelo simple posee un voto y la predicción se realiza por mayoría de votos.
- Bagging: cada modelo simple se entrena con un subconjunto del conjunto de entrenamiento, luego sus predicciones se agrupan por promedio o votación.
- Boosting: los modelos simples estan colocados de modo secuencial de modo que un modelo intenta corregir el error de sus predecesor.

Como se ha dicho, en el caso del clasificador Bagging, éste entrena sus modelos simples con subconjuntos del conjunto de entrenamiento. Existen diferentes modos de elegir dichos subconjuntos, en este caso se hace aleatoramiente a partir del conjunto de entrenamiento extrayendo las muestras con repetición [17].

2.5. Análisis estadístico

Para el evaluar la efectividad del modelo, se recurrirá a la matriz de confusión y los cálculos derivados de esta. Como se puede observar en la siguiente figura, en la diagonal principal de la matriz se encontrarán los sujetos patológicos, tanto aquellos que han sido predichos correctamente (verdaderos positivos), como aquellos que no lo han sido (verdaderos negativos).

REALIDAD \ PREDICCIÓN	
Verdaderos positivos (VP)	Falsos negativos (FN)
Falsos positivos (FP)	Verdaderos negativos (VN)

Figura 3. *Matriz de confusión*

Las métricas que se calcularán a partir de la matriz de confusión serán las siguientes:

- Validez (V): en inglés accuracy, representa la tasa de aciertos con respecto el total. Es una métrica que no se utiliza cuando la sensibilidad y la precisión no tienen la misma importancia en el problema planteado.

$$V = \frac{(VP + VN)}{(VP + FN + FP + VN)}$$

- Sensibilidad (S): en inglés recall, proporción de casos positivos acertados.

$$S = \frac{VP}{(VP + FN)}$$

- Especificidad (E): en inglés specificity, proporción de casos negativos acertados.

$$E = \frac{VN}{(FP + VN)}$$

- Valor predictivo positivo (VPP): en inglés precision, hace alusión a los casos positivos correctamente clasificados entre los que el modelo predice como positivos.

$$VPP = \frac{VP}{(VP + FP)}$$

- F1: es un valor que permite evaluar la precisión y la sensibilidad al mismo tiempo.

$$F1 = 2 * \frac{S * VPP}{(S + VPP)}$$

3. Resultados y discusión

En la siguiente figura, se encuentra la matriz de confusión obtenida tras el entrenamiento y validación del conjunto de datos por validación cruzada:

REALIDAD \ PREDICCIÓN	
845	151
179	481

Figura 4. *Matriz de confusión resultante en el experimento*

En la siguiente tabla se presentan las métricas obtenidas a partir de la matriz de confusión:

V	S	E	VPP	F1
80 %	83 %	75 %	84 %	84 %

Tabla 2. *Análisis estadístico resultante*

Como se puede observar la validez del clasificador Bagging que se ha utilizado en este experimento 80 %, es parecida a la obtenida en [16], donde se obtiene un 82 % de validez. Sin embargo, haría falta más información para poder realizar una comparación válida, puesto que los valores de sensibilidad, especificidad y valor predictivo positivo no están especificados en éste. Además, el conjunto de datos entre estudios no es comparable, en este estudio se está utilizando una sóla tarea vocal en tono neutral, mientras que en [16] se utilizan las tareas vocales /a/, /i/ y /u/ en diferentes tonos.

4. Conclusiones

En este trabajo se ha realizado la evaluación del método de ensemble Bagging para detectar individuos patológicos frente a individuos sanos utilizando el el conjunto de extracción de características automático ComParE. Los resultados de sensibilidad y especificidad son de, 83 % y 75 %, respectivamente. Se obtiene un valor de validez de 80 %, valor similar al obtenido en otros estudios. Para crear el conjunto de datos se han tenido en cuenta las limitaciones de la base de datos Saarbruecken Voice Database y se ha utilizado la tarea vocal /a/ en tono neutral para facilitar la comparativa entre experimentos.

En próximos trabajos, se deberá utilizar otras tareas vocales para ver si existe variabilidad en el rendimiento en función de la tarea y aplicar métodos de selección de características u otros mecanismos para permitir la explicabilidad de los resultados obtenidos que permitiría evaluar qué factores de la voz son los más importantes en voces patológicas, importante para la aplicación clínica final de estos modelos. Así mismo, se realizarán trabajos de identificación de patologías específicas.

Referencias

[1] Cobeta I, Nunez F, Fernandez S. *Patología de la voz*. 2013.

[2] Ziqi F, Wu Y, Zhou C, Zhang X, Tao Z. Class-Imbalanced Voice Pathology Detection and Classification Using Fuzzy Cluster Oversampling Method. *Applied Sciences*, 11(8), 2021.

[3] Wu Y, Zhou CF, Ziqi F, Di W, Zhang X,Tao Z. Investigation and Evaluation of Glottal Flow Waveform for Voice Pathology Detection. *IEEE Access*, 9:30–44, 2021.

[4] Kim H, Jeon J, Han YJ, Joo Y, Lee J, Lee S, Im S. Convolutional neural network classifies pathological voice change in laryngeal cancer with high accuracy. *Journal of Clinical Medicine*, 9(11):1–15, 2020.

[5] Muhammad G, Alsulaiman M, Ali Z, Mesallam TA, Farahat M, Malki KH, Al-nasheri A, Bencherif MA. Voice pathology detection using interlaced derivative pattern on glottal source excitation. *Biomedical Signal Processing and Control*, 31:156–164, 2017.

[6] Pellet P, Phillip E, Mitra S, Holland, TC. Machine Learning-based Voice Assessment for the Detection of Positive and Recovered COVID-19 Patients. *Handbook of Clinical Neurology*, 123(January):45–66, 2020.

[7] Tena A, Claria F, Solsona F, Meister E, Povedano M. Detection of bulbar involvement in patients with amyotrophic lateral sclerosis by machine learning voice analysis: diagnostic decision support development study. *JMIR Medical Informatics*, 9(3), mar 2021.

[8] Smekal Z Harar P, Galaz Z, Alonso-Hernandez JB, Mekyska J, Burget R. Towards robust voice pathology detection: Investigation of supervised deep learning, gradient boosting, and anomaly detection approaches across four databases. *Neural Computing and Applications*, 32(20):15747–15757, oct 2020.

[9] Huckvale M, Buciuleac C. Automated detection of voice disorder in the Saarbrücken voice database: Effects of pathology subset and audio materials. In *Proceedings of the Annual Conference of the International Speech Communication Association*, volume 6, pages 4850–4854. International Speech Communication Association (ISCA), aug 2021.

[10] Massachusetts Eye and Ear Infirmary Voice Disorders Database, 1994.

[11] Barry WJ, Putzer M. Saarbruecken Voice Database, 2008.

[12] Zakariah M, Mohammed R, Alotaibi B, Ajmi-Guo Y, Tran-Trung Y and Elahi K, Mohammad M. An Analytical Study of Speech Pathology Detection Based on MFCC and Deep Neural Networks. *Computational and Mathematical Methods in Medicine*, 2022.

[13] Página web de Linguistic Data Consortium - King Saud University Arabic Speech Database. Accessed on 03.07.2023.

[14] Al-nasheri A, Muhammad G, Alsulaiman M, Ali Z, Mesallam TA, Farahat M, Malki KH, Bencherif MA. An Investigation of Multidimensional Voice Program Parameters in Three Different Databases for Voice Pathology Detection and Classification. *Journal of Voice*, 31(1), 2017.

[15] Eyben F. *Real-time speech and music classification by large audio feature space extraction*. Number 1975. 2014.

[16] Barche P, Gurugubelli K, Kumar-Vuppala A. Towards automatic assessment of voice disorders: A clinical approach. In *Proceedings of the Annual Conference of the International Speech Communication Association*, 2020.

[17] Pagina web documentación Scikit-learn 1.3.0 - Ensemble Methods. Accessed on 03.07.2023.