Uso del PCA como técnica para la identificación de voces

Rodrigo Bisetti
Facultad de Ingeniería
Universidad del Pacífico
Lima, Perú
ra.bisettia@alum.up.edu.pe

Nicolás Mercado
Facultad de Ingeniería
Universidad del Pacífico
Lima, Perú
ra.bisettia@alum.up.edu.pe

Jose Carlos Salinas
Facultad de Ingeniería
Universidad del Pacífico
Lima, Perú
ra.bisettia@alum.up.edu.pe

***Abstract*—La investigación busca comprender el alcance de PCA para la adaptación de voces. Mediante este proceso, se generarán eigenvoices, las cuales concentran las características fundamentales de una voz particular. Se generó el modelo de voz y se evaluó su fiabilidad. Los resultados demostraron que el error relativo para la persona caracterizada con su voz era inferior a la de un audio ajeno, de modo que la adaptación fue efectiva.**

***Index Terms*—PCA, eigenvoices, RMSE, audios, acústica**

I. Introducción

El presente proyecto trata sobre la identificación de voces para un usuario específico. Para esto, se realizará el Análisis de Componentes Principales (PCA). Esta servirá para establecer las características fundamentales del habla de una persona, recopilando audios de entrenamiento para realizar su modelamiento.

Este proyecto es motivado por el interés en la indagación sobre el reconocimiento del habla. Esta es una rama interdisciplinaria de la inteligencia artificial que permite a las computadoras interpretar y comprender el lenguaje humano, constituida en algoritmos y modelos de aprendizaje para convertir señales de audio de habla en vectores con valores numéricos. Particularmente, la exploración sobre la adaptación del habla captura la atención de la investigación. Esta refiere al proceso de ajustar o personalizar los modelos de reconocimiento del habla a las características acústicas de un hablante específico y su entorno acústico.

La constitución de un audio con una voz registrada consta de cuatro variables fundamentales. Las tres primeras son fundamentalmente de configuración humana. La primera, la locución, implica el registro y la proyección de voz de esta, es decir, si su voz es grave o aguda. La segunda ahonda en las características físicas del sonido, la acústica. Esta evalúa las frecuencias altas y bajas, y el volumen. Y la tercera, relacionada más a la cultura, es el lenguaje. Esta moldea cómo una persona expresa determinados fonemas o el ritmo agitado o pausado de su dicción [1]. El cuarto factor constituye todo el entorno acústico donde el audio es grabado, es decir, en todos los factores no humanos de la grabación. Estas constan del ruido en el ambiente o la calidad del micrófono, que pueden alterar la pulcritud con la cual una voz humana se puede grabar. Esta combinación de factores constituye un audio, los cuales serán el input para extraer las cualidades fundamentales de una voz humana.

Dado que los audios guardan, dentro de su espectrograma, todas estas cualidades acústicas a nivel de valores números en vectores, el objetivo es evaluar múltiples audios para extraer las cualidades acústicas principales de un hablante para poder modelar su voz. Los eigenvoices, precisamente, es el modelo de todas las combinaciones lineales que capturan la mayor variabilidad en los datos de una voz. En otras palabras, es el modelo que permite registrar las principales características acústicas de una señal de audio (voz y entorno). Esto permitirá caracterizar a una persona en particular y distinguirla de otra.

Por lo tanto, el presente trabajo cuenta con dos objetivos. En primer lugar, explorar el uso de la técnica PCA como una alternativa para hacer reconocimiento de voz. En segundo lugar, reducir la cantidad de información que debe ser procesada durante el reconocimiento de voz, eligiendo el número óptimo de componentes.

II. Estado del Arte

Existen estudios que ahondan en el uso del PCA para la adaptación de una voz humana. El presente Estado del Arte contrastará los principales trabajos realizados y sus resultados.

Principalmente, el PCA ha sido utilizado en sistemas complejos donde los datos eran redundante. Nittin y Madhusudan [2] demuestra, evaluando 4 frases diferentes de 7 personas, que el PCA es una buena herramienta para reconocer voces y reducir información. No obstante, esta usualmente es acompañada de otros métodos de análisis para alcanzar mayor pulcritud en la labor. En el mismo paper, se utiliza el Independent Component Analysis (ICA), que a partir de la caracterización fundamental, permite separar las voces y clasificarlas. En sí misma, el PCA se limita solo a la caracterización fundamental para los eigenvoices, no a los pasos extra de distinción.

Otro acercamiento al PCA es utilizado por Shabani y Norouzi [3] para distinguir discursos directos con palabras pre-grabadas. En concreto, utilizaron múltiples grabaciones de la misma palabra para reducir la variabilidad de las características acústicas. Utilizaron, además, redes neuronales para la clasificación más efectiva de palabras. Los resultados fueron una reducción considerable en el tiempo de procesamiento

y un menor uso de cantidad de componentes. En este caso, el PCA fue usado para recopilar la mayor variabilidad en la dicción, pero el uso controlado de datos simples simplificó considerablemente el escalamiento del proyecto.

Debido a esto, el presente trabajo resalta la importancia de evaluar los alcances y limitaciones del uso del PCA para la adaptación de la voz con audios en condiciones múltiples.

III. Marco Teorico

A. Conversión de audio a vector

La conversión de un audio a un vector implica la cuantificación de una onda de sonido. Este proceso permite expresar el sonido como un vector. El número de filas dependerá de la frecuencia del audio (medida en kHz) tambien llamado "sample rate". Si la amplitud de la onda ha sido normalizada se escalaron los valores obtenidos de manera que los elementos del vector oscilaran entre [-1,1]. De esta manera el máximo valor que puede tomar la amplitud en 1 mientras que el mínimo vendría a ser -1. Esto se asemeja a la cresta y el valle de una onda. Esto brinda certeza que de que la señal de audio esta representada de manera precisa.

El vector obtenido puede ser graficado. La figura 1 detalla el gráfico del vector para un audio de 5 segundos.

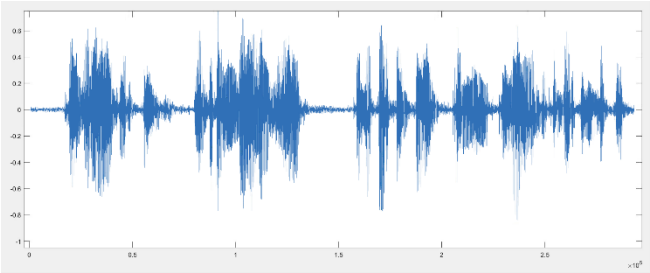

Fig. 1. Gráfico del audio en un vector.

El eje X muestra la cantidad de filas del vector. Este audio de 5 segundos tiene 293,448 filas. Mientras que el eje Y muestra la amplitud de la onda normalizada en ese momento.

Para entrenar nuestro modelo es necesario utilizar 100 distintos audios esto implica construir una matriz de 100 columnas donde cada columna almacena un audio. Esto implica trabajar con una matriz con demasiados elementos de los cuales algunos no tienen un aporte significativo, puesto que sus valores son muy cercano o iguales a 0.

B. Análisis de principales componentes (PCA)

PCA (Principal Components Analysis) es una técnica estadística que transforma un conjunto de datos en componentes principales. Esto permite retener la mayor variación presente de la data set reduciendo las dimensiones de la data. Esta técnica será empleada en este trabajo.

Aplicar PCA nos permitirá reducir la dimensionalidad de la data y únicamente conversar los elementos representativos de todo el set de datos. Esto reduce el tiempo y el espacio necesario para realizar nuestro algoritmo de reconocimiento de voz conservando la mayor variabilidad posible.

A continuación, se muestran las ecuaciones necesarias para realizar PCA. Se emplea la misma nomenclatura que se utilizó en [5].

Siendo X_{ixi} la matriz de variables originales:

$$X = \begin{bmatrix} x_{1,1} & \cdots & \cdots & x_{1,j-1} & x_{1,j} \\ x_{2,1} & \ddots & \cdots & \cdots & x_{2,j} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{i-1,1} & \vdots & \vdots & \ddots & x_{i-1,j} \\ x_{i,1} & x_{i,1} & \cdots & x_{i,j-1} & x_{i,j} \end{bmatrix} \quad (1)$$

Luego, se halló la matriz de covarianzas muestrales (matriz Σ) utilizando la siguiente fórmula.

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \hat{X})(Y_i - \hat{Y})}{n-1} \quad (2)$$

A $X_i - \hat{X}$ se le puede expresar como $X' = \begin{pmatrix} x_1 - \hat{x} \\ x_2 - \hat{x} \\ \vdots \\ x_n - \hat{x} \end{pmatrix}$

A $Y_i - \hat{Y}$ se le puede expresar como $Y' = \begin{pmatrix} y_1 - \hat{y} \\ y_2 - \hat{y} \\ \vdots \\ y_n - \hat{y} \end{pmatrix}$

Esto deja la siguiente fórmula:

$$Cov(X,Y) = \frac{X' \cdot Y'}{n-1} \quad (3)$$

Pero esta fórmula solo aplica cuando se tengan 2 variables. En el caso de múltiples variables, que se encuentren en la matriz R de dimensiones $n * k$ donde n es el numero de observaciones y k el numero de variables. La matriz de covarianzas se puede representar de la siguiente forma:

$$\Sigma = \frac{1}{n-1} R^T R \quad (4)$$

Dado que la matriz $\Sigma - \lambda I$ debe ser singular su determinante debe ser igual a 0. Por ello, resolvemos la ecuación 3 para determinar los autovalores de la matriz V.

$$|(\Sigma - \lambda I)| = 0 \quad (5)$$

La matriz $D_{n \times n}$ (5) incluye los autovalores calculados, a partir de la ecuación 4, en su diagonal. Donde los primeros autovalores retienen la mayor varianza del modelo. Siendo λ_1 el mayor autovalor. El orden de la matriz determina cuántos componentes se retienen.

$$D = \begin{bmatrix} \lambda_1 & 0 & \dots & \dots & 0 \\ 0 & \lambda_2 & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \dots & \vdots \\ \vdots & \vdots & \vdots & \lambda_{n-1} & \vdots \\ 0 & 0 & 0 & 0 & \lambda_n \end{bmatrix} \quad (6)$$

A base de los autovalores calculados previamente, la ecuación 7 nos permite determinar los autovectores.

$$(\Sigma - \lambda I)\,\alpha_i = 0 \tag{7}$$

La matriz U incluye los autovectores asociados a los autovalores. En el mismo orden que el de los autovalores en la matriz.

$$U = \begin{bmatrix} \vdots & \vdots & \dots & \vdots & \vdots \\ \vdots & \vdots & \dots & \vdots & \vdots \\ \alpha_1 & \alpha_2 & \dots & \alpha_3 & \alpha_4 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ \vdots & \vdots & \dots & \vdots & \vdots \end{bmatrix} \tag{8}$$

C. Medición del error

Para el presente trabajo, se utilizó el error cuadrático medio (RMSE), debido a que es menos reactivo antes valores cercanos a 0. Esta observación es importante, pues tal como fue visto en la vectorización del audio, el rango de valores oscila entre [-1;1]. El error es encontrado mediante la siguiente fórmula:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{9}$$

Este error emplea las mismas unidades que la variable en estudio. Esto dificulta su interpretación para el caso de un vector de audio, porque al estar normalizado tiene valores muy cercanos a 0. Por lo tanto, el RMSE tenderá a tener valores muy bajos a pesar de que tengan características muy diferentes.

Debido a esto, se preparó el siguiente cuadro para facilitar la interpretación.

TABLE I
TOLERANCIA PARA EL RMSE

RMSE	***Clasificación***
RMSE >0.02	No es el sujeto
RMSE <0.02	Sí es el sujeto

IV. METODOLOGÍA

A. Recolección de datos de entrenamiento y prueba

Para los datos de entrenamiento, se recopilaron 100 de una misma persona, a quien se denominará "Sujeto 1". En cambio, para los datos de prueba, se seleccionó un audio distinto del "Sujeto 1" y otro de una persona ajena, denominada "Sujeto 2".

B. Establecer los vectores de prueba

Se crean los vectores de prueba tanto para el Sujeto 1 como para el Sujeto 2, a partir de la vectorización de los audios extraídos.

$$X_{Sujeto1} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_i \end{bmatrix}$$

$$X_{Sujeto2} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_i \end{bmatrix}$$

C. Obtención de autovectores y autovalores de la matriz de covarianzas de los datos de entrenamiento

Se crea una matriz X_{train} de tamaño p x n, donde p es la cantidad de audios de entrenamiento y n el total de características del audio. Como se mencionó en el marco teórico, cada columna es un audio distinto.

$$X_{train} = \begin{bmatrix} x_{1,1} & \dots & \dots & x_{1,j-1} & x_{1,j} \\ x_{2,1} & \ddots & \dots & \dots & x_{2,j} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{i-1,1} & \vdots & \vdots & \ddots & x_{1,j} \\ x_{i,1} & x_{i,2} & \dots & x_{i,j-1} & x_{i,100} \end{bmatrix}$$

A continuación, se obtiene la matriz de covarianzas $C_{entrenamiento}$ para los datos de entrenamiento aplicando la fórmula 3 del marco teórico. Luego, se calculan los autovalores $\lambda_{entrenamiento}$ utilizando la fórmula 4 y los autovectores $V_{entrenamiento}$ utilizando la fórmula 7.

$$\lambda_{entrenamiento} = \begin{bmatrix} \lambda_1 & 0 & \dots & \dots & 0 \\ 0 & \lambda_2 & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \dots & \vdots \\ \vdots & \vdots & \vdots & \lambda_{99} & \vdots \\ 0 & 0 & 0 & 0 & \lambda_{100} \end{bmatrix}$$

$$V_{entrenamiento} = \begin{bmatrix} \vdots & \vdots & \dots & \vdots & \vdots \\ \vdots & \vdots & \dots & \vdots & \vdots \\ \alpha_1 & \alpha_2 & \dots & \alpha_{99} & \alpha_{100} \\ \vdots & \vdots & \dots & \vdots & \vdots \\ \vdots & \vdots & \dots & \vdots & \vdots \end{bmatrix}$$

D. Proyección y reconstrucción de los datos de prueba en el espacio de los datos de entrenamiento

Se calcula la matriz de proyección con los datos de prueba.

$$P = X_{prueba} * V_{entrenamiento} \tag{10}$$

En este caso, X_{prueba} es un audio (un vector), el cual puede ser tanto $X_{Sujeto1}$ como $X_{Sujeto2}$. Luego, se realiza la reconstrucción de los audios de prueba en el espacio de los datos de entrenamiento utilizando la fórmula.

$$Reconstrucción = \left(P * (V_{entrenamiento})^T\right) + \mu_{entrenamiento} \tag{11}$$

Donde $\mu_{entrenamiento}$ es la media de los datos de entrenamiento.

E. Cálculo del error utilizando el RMSE para diferentes números de componentes

Se selecciona un rango de valores para el número de componentes, desde 1 hasta el número total de componentes disponibles. Para cada número de componentes k seleccionado, se realiza la proyección y reconstrucción de los datos de prueba utilizando los k primeros autovectores, de manera iterativa. Luego, se calcula el error utilizando el RMSE entre el k-esimo audio reconstruido y el audio original de prueba. Finalmente, se registran los errores obtenidos para cada número de componentes y se generan gráficas o tablas para visualizar el error en función de los componentes retenidos. Se selecciona el número óptimo de componentes con el menor error.

F. Comparación de errores entre Sujeto 1 y Sujeto 2 para verificar la identidad

Utilizando el número óptimo de componentes determinado en el paso anterior, se calcula el error mediante el RMSE para las reconstrucciones de las voces de prueba de Sujeto 1 y Sujeto 2. Se comparan los errores obtenidos para determinar si la voz de prueba corresponde a la persona de los datos de entrenamiento (Sujeto 1) o no. Si el error es significativamente alto, se concluye que el sujeto de prueba no es la persona con la que se creó el modelo.

En cambio, si el error para la voz de prueba de Sujeto 1 es significativamente menor que el error para la voz de prueba de Sujeto 2, se concluye que la persona de la prueba es la misma que la de los datos de entrenamiento. En caso contrario, se concluye que la persona de la prueba (Sujeto 1) es diferente de la persona de los datos de entrenamiento (Sujeto 2).

En resumen, esta metodología permite realizar una comparación de errores entre las voces de prueba de diferentes individuos, lo cual es útil en aplicaciones de autenticación biométrica para determinar la identidad de una persona basada en su voz.

V. Resultados

La figura 2 recopila el RMSE obtenido por cada número de componentes utilizando como audio de prueba al Sujeto 1, es decir, a la persona de la cual ha sido caracterizada su voz. Se observa que el rango del error se reduce de 0.015 a 0.006 a medida que el número de componentes aumenta. Debido a que este rango es menor a 0.02, según nuestra tabla de tolerancia, se identifica que el audio de prueba sí corresponde a la persona entrenada.

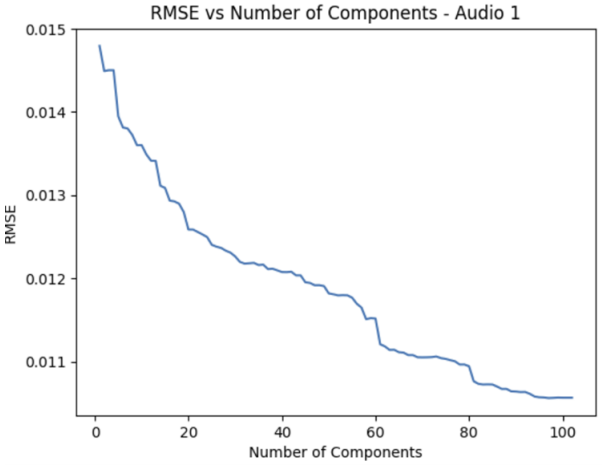

Fig. 2. RMSE obtenido por cada número de componentes usando de prueba al Sujeto 1.

La figura 2, en cambio, presente el RMSE por cada número de componentes iterados, utilizando como audio de prueba al Sujeto 2, es decir, a la persona ajena. Se observa que el rango del error se reduce de 0.1547 a 0.1542 a medida que el número de componentes aumenta. Debido a que este rango es mayor a 0.02, según nuestra tabla de tolerancia, se identifica que el audio de prueba no corresponde a la persona entrenada.

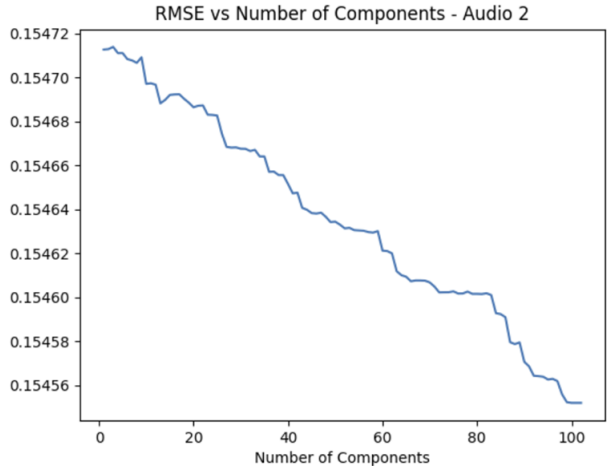

Fig. 3. RMSE obtenido por cada número de componentes usando de prueba al Sujeto 2.

VI. Conclusiones y recomendaciones

Utilizando el criterio establecido en la Tabla I, se determina que el audio 1 pertenece al sujeto uno en cambio el audio 2 pertenece a alguien diferente, aun si el número de componentes aumenta o disminuye. Por lo tanto, se puede aseverar que el uso del PCA para la adaptación de voz sí permite una caracterización que ayude a diferenciar la voz entrenada de una voz ajena. Aunque haya un aumento de componentes, las características acústicas fundamentales de la persona siguen vigentes. La reiteración de audios sirve, más que nada, para

una mejor captación de la variabilidad de la voz humana y su entorno acústico.

Por otro lado, si bien el error esperado para el sujeto 2, debido a la medición numérica de los vectores, debía ser pequeño, es necesario reflexionar sobre qué otras condiciones han propiciado el valor bajo y por qué, a medida que aumentan los componentes, el error disminuye también. Recordando las características que constituyen la grabación de una voz humana en un audio, cuestiones como la distorsión en la calidad del micrófono, los silencios en el uso del lenguaje o las alteraciones locutoras en la voz pueden subvertir la captación de los autovalores. Mientras más audios haya con diferentes frases oracionales, más silencios habrá y más distorsiones captará. De este modo, si bien para una misma persona esto implicaría una caracterización más precisa, otras cualidades más generales también compatibilizarán con la persona ajena. Esto produce, por lo tanto, una mayor compatibilidad entre la voz entrenada y la voz ajena, lo que se demuestra con la disminución del error mientras más componentes hayan.

Esta observación otorga una deducción importante: no siempre más cantidad de dataset implica mayor precisión en los detalles. Lo que propone, en cambio, es el mayor cuidado en el preprocesamiento y normalización de los inputs para una caracterización más significativa. De este modo, el PCA, mediante la medición del RMSE obtenido, permite establecer una métrica para evaluar la cantidad de componentes idóneo para una adecuada caracterización y diferenciación. En este caso, 20 componentes de audio es suficiente para caracterizar adecuadamente a la persona.

Para trabajos futuros, se sugiere la supervisión de ingeniería sonora para obtener audios más limpios y cuyo entorno acústico no distorsione las características acústicas de la persona adaptada. A su vez, se recomienda indagar en las frecuencias altas y bajas de la voz humana para tener una síntesis de voz más precisa.

REFERENCES

[1] Crystal, D. (1941). What is linguistics? Kahle/Austin Foundation. England.

[2] Nitin, K. Madhusudan, R. (2010). Implementation of PCA ICA for Voice recognition and Separation of Speech. Journal on audio, speech and music processing.

[3] Shabani, S. Norouzi, Y. (2016). Speech Recognition Using Principal Components Analysis and Neural Networks. IEEE 8th International Conference on Intelligent Systems.

[4] Saeed, S., Demiroglu, C. King, S. (2017). Using Eigenvoices and Nearest-Neighbours in HMM-Based Cross-Lingual Speaker Adaptation With Limited Data. IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. Vol. 25,

[5] Jolliffe, I. (2002). Principal Component Analysis. Springer Series in Statistics. Springer New York, NY.