Determinación Automática de Calidad de Comunicación en la Operación de Redes Eléctricas

1st Gonzalo Farias
Escuela de Ingeniería Eléctrica
Pontificia Universidad Católica de Valparaíso
Valparaíso, Chile
gonzalo.farias@pucv.cl

2nd Jaime Acevedo
Escuela de Ingeniería Eléctrica
Pontificia Universidad Católica de Valparaíso
Valparaíso, Chile
jaime.acevedo.salinas@gmail.com

3rd Gonzalo Garcia
College of Engineering
Virginia Commonwealth University
Virginia, United States
garciaga3@vcu.edu

4th Ernesto Fabregas
Departamento de Informática y Automática
Universidad Nacional de Educación a Distancia
Madrid, Spain
efabregas@dia.uned.es

5th Gabriel Hermosilla
Escuela de Ingeniería Eléctrica
Pontificia Universidad Católica de Valparaíso
Valparaíso, Chile
gabriel.hermosilla@pucv.cl

6th Sebastián Dormido-Canto
Departamento de Informática y Automática
Universidad Nacional de Educación a Distancia
Madrid, Spain
sebas@dia.uned.es

***Resumen*—The quality of verbal communication between an operator and the control center, understood as the absence of uncertainty, is a key factor in critical processes such as the restoration of service in electrical networks. In the present work, an uncertainty classifier is proposed based on the analysis of formants obtained through the spectrogram of the audio signal that collects the communication between an operator and a supervisor. By generating a feature vector based on clustering samples of adjacent similar formants, the SVM classifier is executed. A classifier based on the word repetition rate is also applied as sentiment analysis on the transcribed text. To determine the best performance of the model, classification metrics are used, prioritizing avoiding losing recordings with uncertainty (false negatives). Finally, the parameters that determine the best performance of the model are defined.**

***Index Terms*—verbal communication uncertainty, spectral analysis, speech analysis, Support Vector Machine (SVM)**

I. Introducción

La gestión de tiempo real de redes eléctricas es un proceso de misión crítica [1]–[3], por eso continuamente se deben hacer esfuerzos por eliminar las fuentes o inductores de error, particularmente en la comunicación entre el Centro de Control y el personal de operación dispuesto en terreno [4].

Para identificar estos inductores de error se proponen dos métodos complementarios. El primero, denominado análisis de formantes, se basa en extraer del espectrograma el contenido de frecuencia relevante, los dos primeros formantes, y determinar el grado de incertidumbre en función de cómo estos formantes cambian a lo largo del tiempo (ver [5], [6] para trabajos relacionados con contenidos de audiofrecuencia). El segundo método, llamado *Speech-to-text* [7], [8] (voz a texto), consiste básicamente en inferir la existencia de incertidumbre en el orden de las palabras utilizadas en el audio, lo que refleja el sentimiento de los hablantes; consulte [9] para conocer un enfoque estrechamente relacionado.

El primer método propone diseñar un clasificador de la incertidumbre transmitida en la comunicación verbal mediante la identificación de fonemas presentes en situaciones de incertidumbre (por ejemplo: 'eeeeeee....', o 'mmmmmmm...'). Estas situaciones son identificadas a través del análisis de los formantes, que corresponden a la frecuencia de mayor magnitud para cada instante de tiempo de la comunicación. Los fonemas del habla quedan descritos por el primer y segundo formante, que se obtienen mediante el análisis del espectrograma del archivo de audio de la grabación de la comunicación. A través de un proceso de análisis de los formantes se procede a generar un Vector de Características, y luego a utilizar el algoritmo SVM (Support Vector Machine) [10] para obtener un clasificador de incertidumbre. Se realizan diversas variaciones de parámetros con el objetivo de medir el desempeño del SVM mediante métricas propias de reconocimiento de patrones, y clasificación de objetos.

El segundo enfoque, basado principalmente en detectar el contenido oculto de sentimiento o emoción, en este caso asociado a vacilaciones o dudas, utiliza el orden en que se incluyen las palabras en el texto del audio. Los trabajos relacionados incluyen [11], [12]. Este enfoque se basa en una técnica llamada *Speech-to-text*. En este caso es importante

considerar que se debe utilizar un proceso entrenado para el idioma español, idealmente latinoamericano o chileno, por lo que se utilizó una herramienta especial para realizar esta conversión de audio a texto. Luego se construyó una bolsa de palabras a partir de todas las palabras existentes en el audio y, en consecuencia, se realizó un proceso de conversión de palabras a números para un análisis numérico efectivo utilizando SVM. En este segundo método se emplean métricas de rendimiento similares.

Para ambos métodos, se identifican los parámetros que generan el mejor desempeño del SVM para el caso estudiado, cumpliendo con el objetivo de detectar comunicaciones que tengan presente niveles de incertidumbre que requieran revisión por parte de un supervisor, dentro de un contexto de mejora continua del proceso.

II. Materiales y Métodos

II-A. Descripción e Importancia del Problema

En el ámbito de la gestión de redes eléctricas, un Centro de Control es la unidad organizacional responsable de la supervisión y coordinación de la operación en tiempo real del sistema eléctrico, y tiene por función asegurar la continuidad de suministro y la seguridad de ese sistema, sus activos y las personas que intervienen en él, manteniendo los parámetros técnicos dentro de los rangos legales definidos [13].

Para cumplir el objetivo del Centro de Control existe el rol de Operador, que es la persona encargada de monitorear y telecomandar el sistema eléctrico a través de equipamiento en terreno, telecomunicaciones y SCADA [14]. En caso de necesitar la realización de labores presenciales en terreno, es el responsable de instruir remotamente las maniobras necesarias para realizar los trabajos en forma segura.

Es por lo anterior que existen estrictos protocolos de comunicación entre el Operador y el personal de terreno, de forma que las instrucciones emitidas sean efectivamente comprendidas y adecuadamente ejecutadas. No obstante la existencia de estas definiciones, la complejidad de las comunicaciones verbales genera ciertos riesgos cuando no hay un buen entendimiento entre los interlocutores. Esta complejidad está dada, entre otros, por los trabajos a realizar, los estados de ánimo, el contexto, y la calidad técnica del canal de comunicación y sus equipos terminales. Para evaluar los resultados en la práctica, se realizan revisiones aleatorias a las grabaciones de estas interacciones verbales. Este método, si bien es efectivo para las grabaciones revisadas, tiene las siguientes desventajas:

- Alto costo en tiempo, pues escuchar las grabaciones toma al menos el mismo tiempo que la comunicación original.
- Baja cobertura, pues se logra revisar sólo un subconjunto menor de grabaciones, perdiendo así la oportunidad de capturar una mayor cantidad de situaciones a mejorar.

Para superar las desventajas del método actual, y de esa forma aumentar la efectividad del proceso de control de calidad, se propone la utilización de un proceso automático de análisis de grabaciones que permita:

- Revisar el $100\,\%$ de las grabaciones.
- Clasificar comunicaciones que generan riesgo de error por falta de claridad.

La implementación se basará en algoritmos de Inteligencia Artificial ejecutados en modo batch (off line), que procesarán archivos de audio con grabaciones de los diálogos con instrucciones. Mediante extracción de características y utilización de SVM se clasificará la calidad de la interacción entre el Centro de Control y el operador de terreno. La operación de tiempo real de un sistema eléctrico califica dentro de los procesos de misión crítica. Por este motivo, cualquier error o malentendido puede generar graves consecuencias indeseadas, tanto en la continuidad y calidad del suministro, como en el estado de los activos, la salud de los trabajadores, y finalmente tener impacto en la población y en la reputación de la empresa. Es por ello que se deben reducir al máximo los inductores de eventos no deseados, siendo la comunicación verbal entre Operador y personal de terreno uno de los aspectos más sensibles, dada la interacción directa entre dos personas [15], [16].

Desde una mirada preventiva y anticipativa, es de interés identificar comunicaciones reales que generen riesgo de error en ejecución de operaciones manuales en terreno, para realizar las intervenciones y retroalimentaciones que permitan ir eliminando estos riesgos de malentendidos o de instrucciones confusas o erróneas. El beneficio directo se traduce en reducción de tiempo y aumento de la eficiencia y eficacia de la revisión de los archivos de audio, ya que el clasificador diseñado debe eliminar aquellas conversaciones sin incertidumbre o con baja probabilidad de incertidumbre, seleccionado y ordenando aquellos audios que tienen mayor incertidumbre detectada.

II-B. Estrategias de Solución

El objetivo de este trabajo, basado en parte en [17], es identificar las grabaciones de audio que tienen mayor probabilidad de contener comunicaciones con incertidumbre. Por lo tanto, el problema a resolver se debe centrar en la siguiente pregunta: ¿Cómo identificar la incertidumbre de una comunicación verbal grabada en un archivo de audio?

La pregunta anterior tiene, al menos, dos posibles formas de implementar una respuesta:

- Identificación de palabras o conjunto de palabras que representen incertidumbre en la comunicación, mediante la definición de un conjunto de palabras que no se deberían usar o repetir en un contexto de instrucciones de una misión crítica. Por ejemplo: ‘no’, ‘no sé’, ‘no lo sé’, ‘espera’, ‘voy a averiguar’, ‘no te entiendo’, ‘¿estás seguro?, u otras.
- Identificación de fonemas que representan incertidumbre en la comunicación. Por ejemplo, ‘aaaaaa’, ‘eeeeeee’, ‘mmmmmmm’.

*II-B1. Enfoque de Análisis de "*Speech-to-text*":* La implementación de la primera estrategia se realiza mediante un proceso de "*Speech-to-text*". En este caso es importante considerar que se debe utilizar un proceso entrenado para el idioma español, idealmente latinoamericano o chileno. Para realizar este proceso de transformación de audio a texto se

utilizó Whisper [18], desarrollado por openAI. Este sistema fue entrenado con una variedad de idiomas, uno de ellos es el español, por lo que se decidió ocupar esta herramienta para la conversión. Una vez obtenido el texto escrito, se debe realizar la identificación del uso de palabras o frases contenidas en un listado de palabras que generan incertidumbre.

Existe cierta limitación en la creación de una bolsa de palabras características mencionadas anteriormente, que es no mencionar ciertas palabras o contextos en los cuales puede haber incertidumbre, por lo que se utilizará el método NLP (Natural Language Processing) para la extracción de características. En este método se asigna un valor numérico a cada palabra presente en los textos para formar un diccionario con estas, esto es conocido como tokenizar, luego de esto se reescriben los textos como una secuencia de números en base a las palabras existentes dentro de éste. El objetivo principal de este método es obtener las características según el orden y repetición de estas palabras más que la palabra en sí. Así, se puede obtener de manera más precisa lo que se quiere expresar, es decir, el "sentimiento" de la oración.

- Tokenización: El proceso basado en un subcampo de NLP, consiste en convertir las palabras en un número único correspondiente como identificador. Se construye una bolsa de palabras asignando estos números a cada palabra sin repetirlas. Una vez construido el diccionario, se convierten los textos en vectores numéricos: VN^i_{SVM}, $i = \{1 \dots m\}$, como se muestra en la Fig. 1.

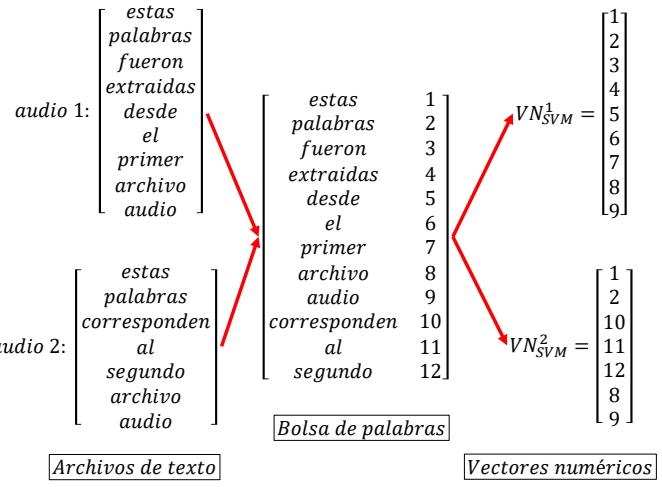

Figura 1. Proceso de Tokenización para dos textos arbitrarios.

Dos preocupaciones principales, debidas a la naturaleza variable de los datos, son el redimensionamiento de todos los vectores añadiendo ceros, e igualando sus tamaños, lo que se denomina relleno cero. Y la segunda preocupación, corresponde a la bolsa de palabras que debe ajustarse incorporando nuevas palabras no vistas durante el entrenamiento.

Del proceso de tokenización de palabras se obtienen los textos con números según las palabras contenidas como se menciona anteriormente, a modo de ejemplo se muestra el resultante de un par de oraciones antes y después de su tokenización:

1. Roberto, no me quedó claro.
2. Roberto, me quedó claro.

Se asignan valores numéricos a cada palabras:

1. $[1, 2, 3, 4, 5]$.
2. $[1, 3, 4, 5]$.

La única diferencia es la palabra "no" pero el resto de la estructura es igual, de esta manera verificando el orden de las palabras se puede extraer el contexto y su significado a través de entrenamiento por redes neuronales.

Para garantizar una entrega optima de los datos, se lleva a cabo un proceso relleno de ceros de las secuencias de manera que todas queden de la misma longitud, esto se realiza agregando 0 hasta que todas las secuencias queden del largo deseado:

1. $[1, 2, 3, 4, 5]$.
2. $[1, 3, 4, 5, 0]$.

En la fase de prueba para evitar errores con palabras nuevas que no estén previamente en el diccionario, se asigna un valor numérico estándar, de esta manera la oración mantiene su longitud y estructura original y se evita la pérdida de información.

- Análisis basado en SVM: Los datos recopilados mediante el enfoque de voz a texto se utilizan con una SVM. En este caso, los vectores característicos a inyectar en el clasificador SVM incluyen la frecuencia de repetición relevante de cada palabra dentro de la totalidad del audio, frecuencias que han superado un umbral ajustado. El primer filtrado consiste en descartar palabras de la bolsa de palabras, con 3 o menos letras, en un intento de limpiar los datos de palabras con funciones más relacionadas con conectar ideas u oraciones, como la mayoría de las preposiciones. Con el objetivo de concentrar las características más importantes contenidas en las frecuencias de repetición y reducir las dimensiones generales de los datos, estos vectores numéricos se someten a una técnica llamada Factorización Matricial No Negativa [19]. Estos nuevos vectores numéricos, VF^i_{SVM}, $i = \{1 \dots m\}$, luego se introducen en el motor SVM. De manera similar, para el entrenamiento se empleó una función Kernel lineal, con un costo de factor de penalización, ver Fig. 2.
- Filtraje de datos por frecuencia de repetición: Aquí se utiliza la bolsa de palabras ya calculada, así como la tokenización. La diferencia es que en lugar de cambiar su tamaño mediante el método de incrustación, los datos se asignan a sus frecuencias de repetición. Se calcula la frecuencia de repetición acumulada por palabra dentro de todos los audios. Luego se define un umbral para seleccionar aquellas frecuencias por encima de él, y sus palabras asociadas, que potencialmente conllevan un sentimiento inherente asociado con el nivel de incertidumbre.

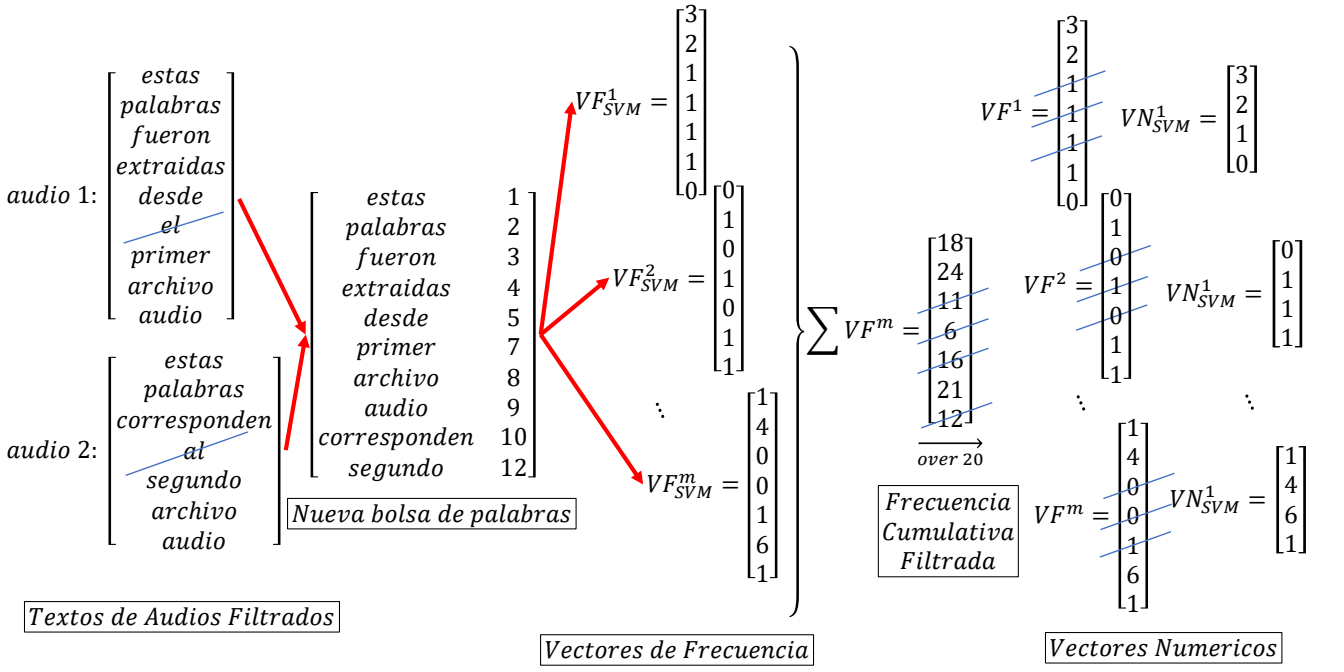

Figura 2. Creación del vector de frecuencia de dos textos arbitrarios.

- Factorización matricial no negativa (NNMF) y validación cruzada de *K-fold*: Esta técnica (ver [20] para más detalles), es un análisis multivariado de álgebra lineal, donde una matriz se factoriza en dos matrices, en dimensionalidad reducida, ayudando al trabajo con grandes volúmenes de datos. Los datos se dividen aleatoriamente en subconjuntos de igual tamaño. De los subconjuntos, un subconjunto único se mantiene fuera del proceso de entrenamiento, pero se utiliza más adelante para realizar pruebas. Este proceso se repite el mismo número que la partición, intercambiando el subconjunto elegido para probar, por lo que se consideran todos. Luego se promedian los resultados (ver [21]).

II-B2. Identificación de los formantes: La segunda estrategia se realiza mediante un proceso más complejo, basado en la identificación de formantes. Los formantes corresponden a las frecuencias de mayor intensidad presentes en una señal de audio. Los formantes se obtienen mediante un análisis en el dominio de la frecuencia a través de un espectrograma que representa la intensidad de la señal para cada frecuencia dentro de su ancho de banda, y para cada instante de tiempo discreto de la señal grabada. En idioma español, los fonemas quedan determinados por la frecuencia y magnitud de los 2 primeros formantes. De esta forma existen las denominadas Cartas de Formantes, que permiten identificar el uso de las vocales, ver Fig. 3.

Si bien la utilización de la Carta de Formantes permite identificar el uso de vocales en idioma español, no considera el uso de consonantes que habitualmente se utilizan para encubrir la incertidumbre. Por ejemplo, el uso de 'mmmmmm' mientras se busca una respuesta. Dada la desventaja anterior, más

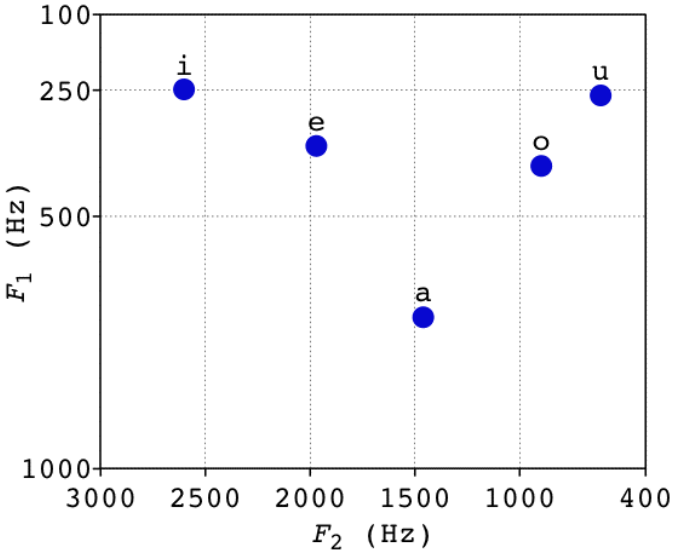

Figura 3. Carta de Formantes para Vocales en Español.

que buscar vocales mediante Carta de Formantes, se buscará identificar tramos de audio que tengan similares formantes contiguos, ya que eso representa el uso del mismo fonema. Esta estrategia está basada en que no es de mayor interés identificar cuál es el fonema utilizado, sino que interesa identificar un fonema que se usa durante una cierta ventana de tiempo, que una vez superada en duración, es señal de incertidumbre en la comunicación verbal.

El proceso de implementación de esta estrategia se muestra en Fig. 4.

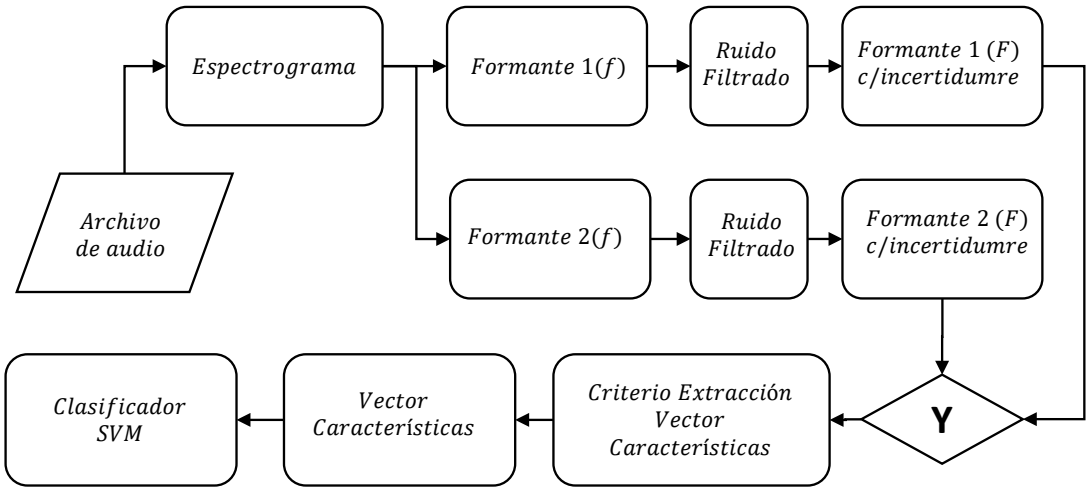

Figura 4. Proceso de Generación del Clasificador de Incertidumbre.

- Criterios de Extracción de Características:

Del proceso de detección de formantes se extrae una matriz con el siguiente formato:

$$M_F = [t, f_{f1}, m_{f1}, f_{f2}, m_{f2}] \tag{1}$$

donde:

- M_F: matriz de formantes, de dimensión [cantidad de muestras del archivo de audio; 5].
- t: tiempo, que va desde 0 hasta la duración del archivo de audio.
- f_{f1}: frecuencia del formante 1.
- m_{f1}: magnitud del formante 1.
- f_{f2}: frecuencia del formante 2.
- m_{f2}: magnitud del formante 2.

De acuerdo a lo indicado anteriormente, interesa detectar secciones de audio donde el fonema se mantenga constante. Es decir, ventanas de audio que cumplan con las siguientes condiciones:

$$\begin{aligned} f_{f1}^{n+1} - f_{f1}^{n} < tol \\ f_{f2}^{n+1} - f_{f2}^{n} < tol \end{aligned} \tag{2}$$

con n n-ésima muestra del archivo de audio de tiempo discreto. El parámetro tol se utiliza para detectar frecuencias muy similares, pero no necesariamente idénticas, dadas las leves variaciones tonales de la voz para la misma sílaba.

Mediante el uso de un archivo de audio de referencia, que contiene las 5 vocales, se determina empíricamente que un valor adecuado que permite distinguir fonemas contiguos parecidos es 50 [Hz]. Valores menores generan un modelo de baja sensibilidad, mientras que valores mayores tienden a clasificar como incertidumbre a fonemas que no están relacionados.

El Vector de Características que se utilizará está definido por la duración de las secciones de fonemas similares, como interpretación de incertidumbre. De esta forma, el vector queda definido por:

- T1: tiempo acumulado de fonemas similares cuya duración está contenida entre 200 y 300 [ms].
- T2: tiempo acumulado de fonemas similares cuya duración está contenida entre 300 y 500 [ms].
- T3: tiempo acumulado de fonemas similares cuya duración está mayor a 500 [ms].
- T4: porcentaje de tiempo con incertidumbre (T1+T2+T3) respecto de la duración total del audio.

A modo ilustrativo, se incorpora el resultado del algoritmo de extracción de características implementado en MATLAB R2022a [22], ver Fig. 5. En este caso, el Vector de Características del archivo de audio es $Vc = (0; 0{,}29434; 1{,}3069; 26{,}1036)$.

III. Resultados y Discusión

III-A. Entrenamiento y Validación

Para entrenar los algoritmos hay un conjunto de 190 grabaciones de audio etiquetadas, incluidas 31 etiquetadas como "con incertidumbre" y 159 etiquetadas como "sin incertidumbre". Con ese universo de grabaciones etiquetadas, y dado el desequilibrio en el número de audios con y sin incertidumbres, se prueban diferentes proporciones de archivos de cada clase para evaluar el efecto de este desequilibrio. Estas proporciones se muestran en la Tab. I.

Tabla I
Cinco diferentes proporciones de audios "con incertidumre" del total del dataset.

Con incertidumbre	Sin incertidumbre	Porcentaje
31	159	$\approx 16\,\%$ (31 de 190)
31	66	$\approx 32\,\%$ (31 de 97)
31	33	$\approx 48\,\%$ (31 de 64)
31	18	$\approx 64\,\%$ (31 de 49)
31	8	$\approx 80\,\%$ (31 de 39)

Para obtener resultados estadísticamente más representativos, se realizaron ejecuciones de 100 para cada una de las combinaciones que se muestran en la Tab. I, seleccionando

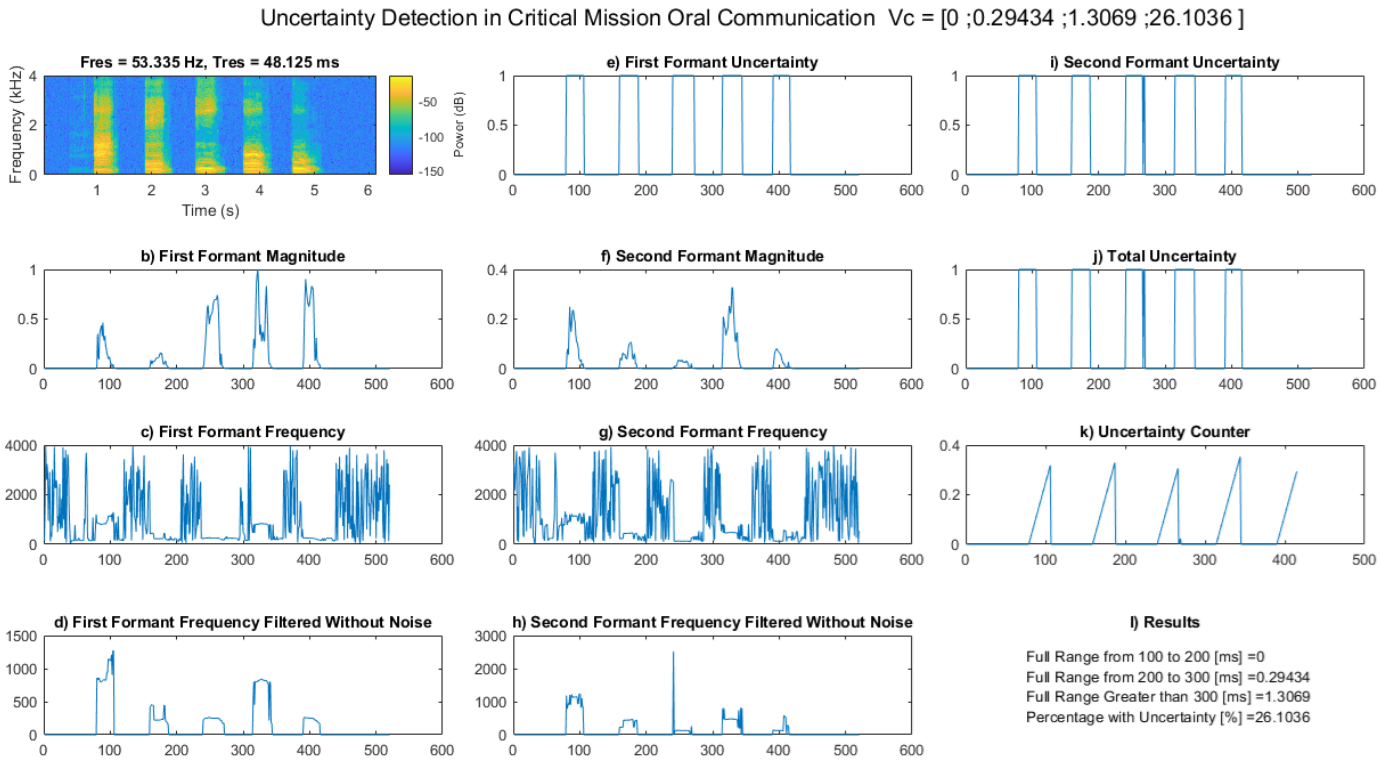

Figura 5. Resultados del Proceso de Extracción de Características.

aleatoriamente el subconjunto sin incertidumbre entre las 159. Los resultados se promediaron para cada categoría, por lo que se presentan valores medios.

Dada la naturaleza desequilibrada de los datos disponibles, se probaron diferentes proporciones de audio para obtener una idea del efecto de esta característica. En teoría, y así se desprende de los resultados, se obtenían mejores resultados cuando las proporciones eran más equivalentes. Este fue el caso en ambos métodos.

Se diseñaron pruebas que incluyeron los 31 audios con incertidumbre y un número variable de audios sin incertidumbre (ver Tab. I), correspondientes a los porcentajes: $16\,\%$, $32\,\%$, $48\,\%$, $64\,\%$, $80\,\%$, correspondiendo el $16\,\%$ al caso donde se utilizaron todos los 159 audios sin incertidumbre. Para cada uno de estos casos se ejecutaron hasta 100 épocas, barajando aleatoriamente los audios sin incertidumbre de ser incluidos. Se promediaron las métricas resultantes, Precision P, Recall R, Accuracy A y $F_1 - score$, para cada caso.

Para cada caso, la validación se realizó después del entrenamiento. Del total de audios, estos son los 31 con incertidumbre más el número particular utilizado de los sin incertidumbre, se hizo una partición seleccionada al azar para reservar una fracción para el entrenamiento real y el resto para validación. Es decir, una vez finalizado el entrenamiento, se presentan al algoritmo nuevos datos, no utilizados durante el entrenamiento, pero que pertenecen al caso actual.

Para cada uno de los casos mostrados en la Tab. I, y cada una de las 100 ejecuciones, se configuraron las siguientes particiones para dividir los audios para entrenamiento y validación: $10\,\%$, $20\,\%$, $30\,\%$, $40\,\%$, $50\,\%$, $60\,\%$, $70\,\%$, $80\,\%$, $90\,\%$, donde el porcentaje indica el número de audios seleccionados para validación del total.

III-B. Análisis de Resultados

III-B1. Formantes: De acuerdo a las corridas resultantes de la variación de parámetros indicada anteriormente, las curvas de desempeño de Precision, Recall, Accuracy y $F_1 - score$ se muestran en la Fig. 6.

Del análisis de las curvas de desempeño de todas las métricas se puede observar lo siguiente:

- Para los casos con proporciones bajas de "con incertidumbre" ($16\,\%$ y $32\,\%$), el rendimiento es en general muy bajo.
- Para los casos con proporciones intermedias ($48\,\%$ y $64\,\%$), el rendimiento es mejor y más estable para diferentes rangos de validación en comparación con los demás.
- La proporción con $80\,\%$ parece ser el mejor resultado para el método de Análisis de Formantes. La Tab. II detalla el rendimiento de la métrica para $80\,\%$.

III-B2. Speech-to-text: Este análisis se basa en la frecuencia de repetición de palabras en cada texto. La Fig. 7 muestra las frecuencias para ambos conjuntos de audios juntos (desde la palabra 250 hasta la palabra 450, del total de palabras de más de 4 mil) asociadas con el subconjunto de palabras en la bolsa de palabras. Se ha establecido un umbral de 20 repeticiones para reducir las palabras más relevantes en función de su prevalencia.

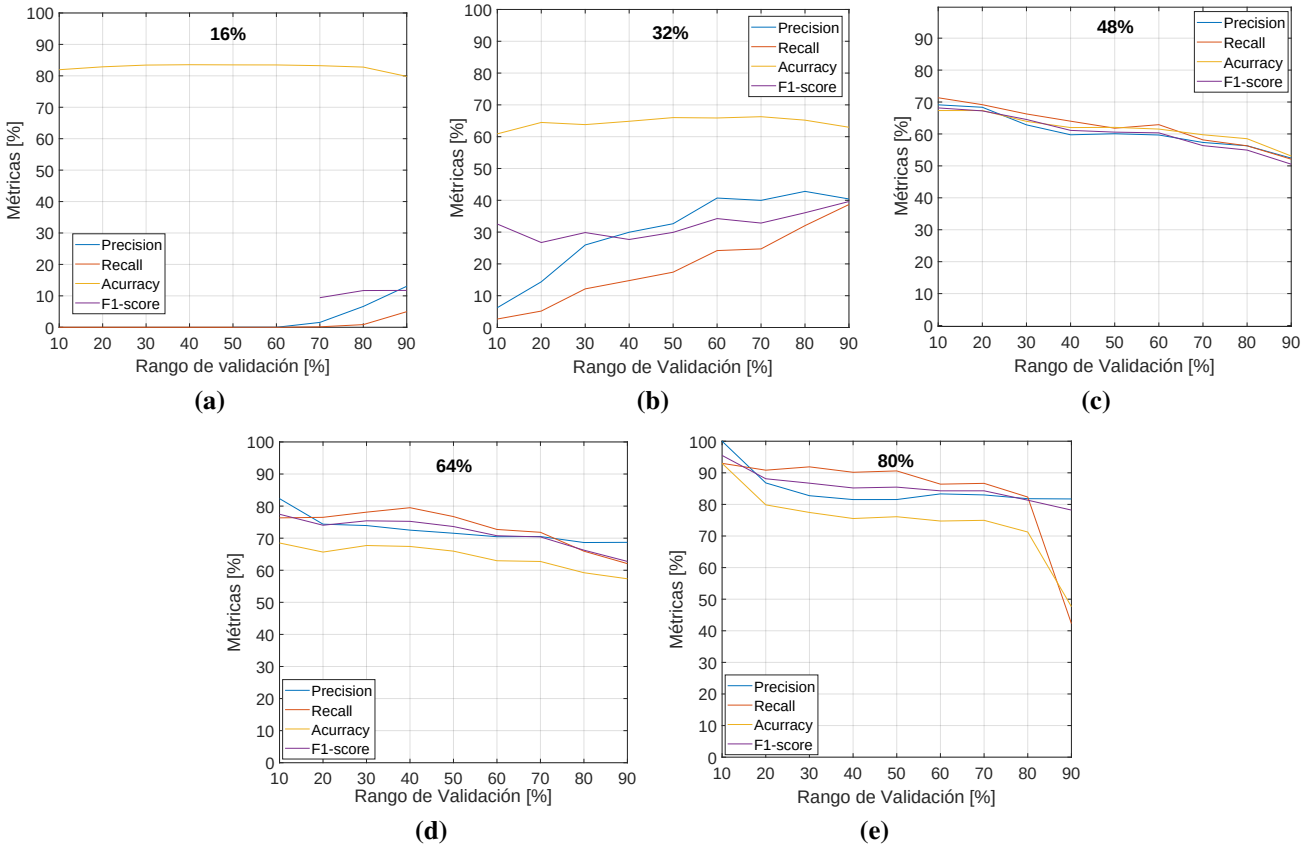

Figura 6. Métricas de Análisis de Formantes para Proporciones mostradas en la Tab. I: **(a)** 16 %, **(b)** 32 %, **(c)** 48 %, **(d)** 64 %, **(e)** 80 %.

Tabla II
MÉTRICAS DE ANÁLISIS DE FORMANTES PARA UNA PROPORCIÓN DE 80 % (VER TABLA. I).

Validación (%)	Precision	Recall	Accuracy	F1
10	1.0000	0.9300	0.9300	0.9550
20	0.8680	0.9083	0.7986	0.8812
30	0.8276	0.9189	0.7745	0.8672
40	0.8153	0.9017	0.7553	0.8521
50	0.8154	0.9060	0.7611	0.8547
60	0.8333	0.8641	0.7474	0.8430
70	0.8300	0.8666	0.7496	0.8430
80	0.8181	0.8233	0.7126	0.8134
90	0.8173	0.4229	0.4766	0.7820

Como antes, los datos se analizan con base en la Tab. I, y luego se aplica una factorización matricial no negativa (NNMF, [23]) con 15 características destacadas. Luego, cada resultado se multiplica por 10 para su clasificación. De manera similar, las curvas de rendimiento de Precision, Recal, Accuracy y $F_1 - score$ se muestran en la Fig. 8. Se pueden enunciar los siguientes puntos:

- En un patrón similar, el rendimiento es en general muy bajo, para proporciones bajas de audios "con incertidumbre" (16 % y 32 %).
- El rendimiento mejora relativamente para los casos intermedios (48 % y 64 %), en función de los rangos de validación.

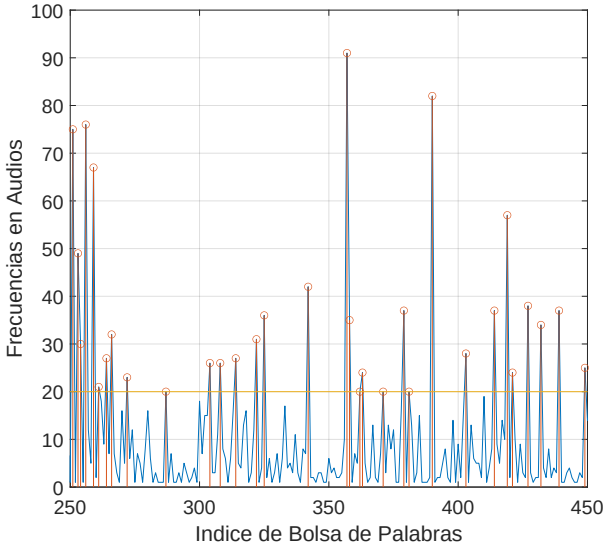

Figura 7. Frecuencia de Repetición de Palabras en cada Texto.

- Como confirmación de resultados anteriores, la proporción del 80 % da mejores resultados en general. La Tab. III detalla el rendimiento de la métrica para 80 %.

III-B3. Comparación de Métodos Propuestos: Tab. IV muestra una comparación entre los tres enfoques desarrollados

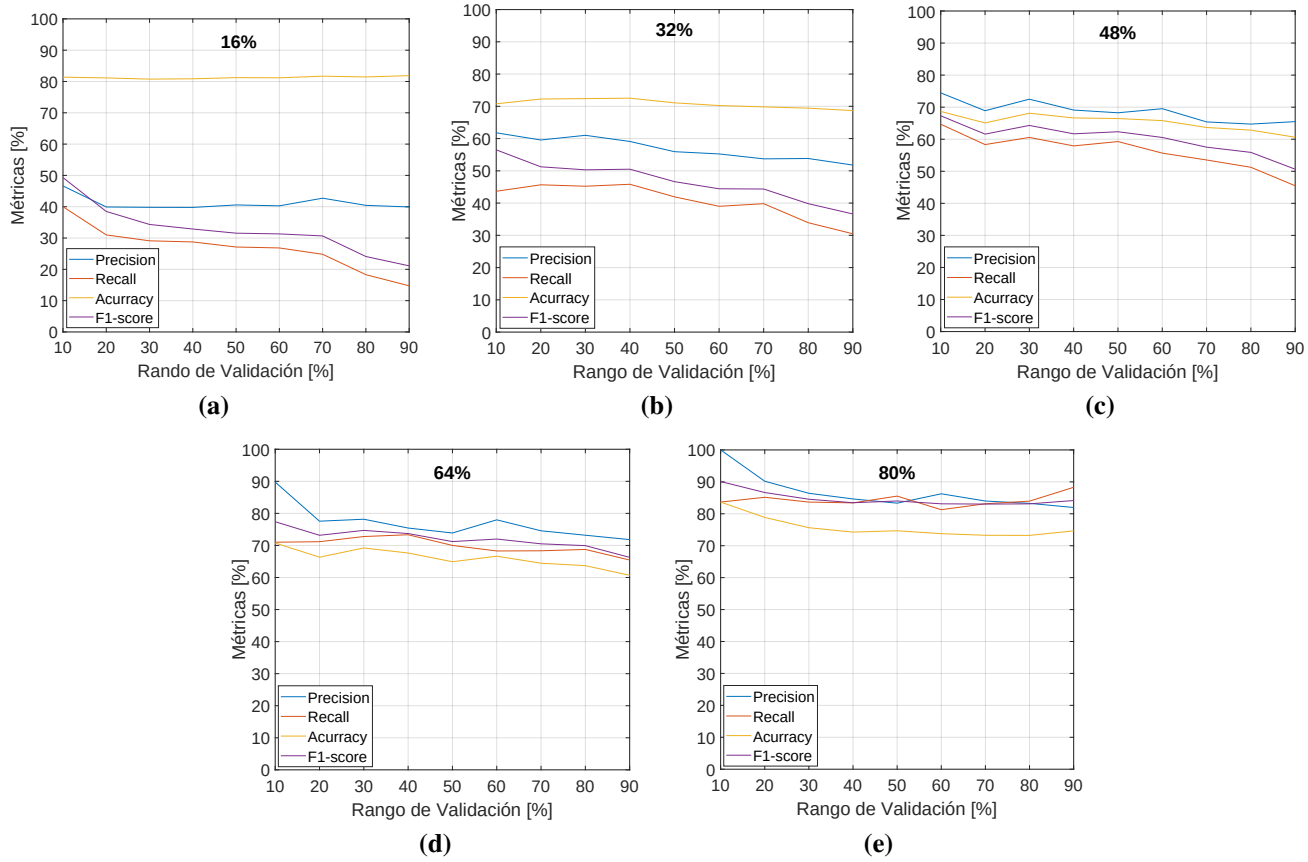

Figura 8. Análisis de Métricas de *Speech-to-text* SVM: **(a)** 16 %, **(b)** 32 %, **(c)** 48 %, **(d)** 64 %, **(e)** 80 %.

Tabla III
MÉTRICAS PARA SPEECH-TO-TEXT SVM, BASADO EN FACTORIZACIÓN DE MATRICES NO NEGATIVAS. PROPORCIÓN DEL 80 % (VER TABLA. I).

Validation (%)	Precision	Recall	Accuracy	F1
10	1.0000	0.8366	0.8366	0.9010
20	0.9019	0.8516	0.7885	0.8666
30	0.8641	0.8366	0.7563	0.8456
40	0.8464	0.8345	0.7426	0.8345
50	0.8332	0.8553	0.7468	0.8400
60	0.8626	0.8127	0.7377	0.8312
70	0.8399	0.8303	0.7326	0.8303
80	0.8310	0.8395	0.7323	0.8310
90	0.8197	0.8829	0.7464	0.8414

en este trabajo. Para comparar el rendimiento, hemos seleccionado la puntuación $F_1 - score$, que equilibra las métricas de precisión y recuperación. Se puede observar que se obtienen mejores resultados cuando la proporción es 80 %, como se mencionó anteriormente. El método con mayor $F_1 - score$ corresponde al Formante con 95,5 %.

De acuerdo a lo que se muestra en la Tab. IV, para el caso en que la mayoría de los audios utilizados para la validación (16 % y 32 %) no tengan incertidumbre, el método *Speech-to-text*, funciona mejor. Si bien se da el caso en el que la mayoría de los datos de validación tienen incertidumbre (64 % y 80 %), el método que mejor funciona es el Formante.

Tabla IV
COMPARACIÓN DE MÉTRICA F_1 PARA AMBOS ENFOQUES. SE MUESTRAN VALORES MÁXIMOS.

Approach	16 %	32 %	48 %	64 %	80 %
1 Formant (SVM)	11.6	39.6	68.1	77.4	**95.5**
2 *Speech-to-text* (SVM)	**49.2**	**56.5**	67.3	77.3	90.1

IV. CONCLUSIONES

En este artículo, mostramos el desarrollo de dos métodos para la clasificación automática de la incertidumbre en las comunicaciones verbales. Los resultados demuestran que la detección y clasificación de la incertidumbre es factible, a pesar del reducido conjunto de datos disponibles.

El primer método, llamado análisis de formantes, extrae del espectrograma el contenido de frecuencia relevante de los dos primeros formantes, por tiempo de muestra, y determina el nivel de incertidumbre dentro del audio, en función del cambio que estos formantes sufren con el tiempo. El segundo método, llamado *Speech-to-text*, infiere del orden de las palabras utilizadas en el audio, el nivel de incertidumbre, interpretando el sentimiento de los hablantes. Todos los resultados, para la mayoría de las métricas de rendimiento, muestran que la proporción del 80 %, que corresponde a 31 audios con incertidumbre y 8 sin incertidumbre, parece proporcionar el mejor resultado. Como limitaciones a lo propuesto, está presente la

dificultad para obtener nuevos datos, afectando el rendimiento de los modelos basados en datos. Y debido a la naturaleza inherente de este tipo de comunicaciones, los datos tienden a estar desequilibrados. Trabajos futuros deberían considerar la recopilación de un conjunto más grande de datos para mejorar el equilibrio entre los dos tipos de audio.

REFERENCIAS

[1] G. Biard and G. A. Nour, "Industry 4.0 contribution to asset management in the electrical industry," *Sustainability*, vol. 13, no. 18, p. 10369, 2021.

[2] V. Listyuhin, E. Pecherskaya, O. Timokhina, and V. Smogunov, "System for monitoring the parameters of overhead power lines," in *Journal of Physics: Conference Series*, vol. 2086. IOP Publishing, 2021, p. 012059.

[3] K. Schröder, G. Farias, S. Dormido-Canto, and E. Fabregas, "Comparative analysis of deep learning methods for fault avoidance and predicting demand in electrical distribution," *Energies*, vol. 17, no. 11, 2024. [Online]. Available: https://www.mdpi.com/1996-1073/17/11/2709

[4] S. A. Melin, "Lightning location system increases personnel safety in swedish power transmission network," in *3D Africon Conference. Africon'92 Proceedings (Cat. No. 92CH3215)*. IEEE, 1992, pp. 497–500.

[5] Y. T. Suy, "Information content of a sound spectrogram," *J. Audio Eng. Soc*, vol. 15, no. 4, pp. 407–413, 1967. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=1079

[6] A. Uğur, G. Hüseyin, K. Faheem, A. Naveed, W. Taegkeun, and B. Abdusalomov, "Automatic speaker recognition using mel-frequency cepstral coefficients through machine learning," *Computers, Materials & Continua*, vol. 71, pp. 5511–5521, 01 2022.

[7] T. Ayushi, P. Navya, S. Pinal, S. Simran, and A. Supriya, "Speech to text and text to speech recognition systems-areview," *IOSR J. Comput. Eng*, vol. 20, no. 2, pp. 36–43, 2018.

[8] R. B. Raghavendhar and M. Erukala, "Speech to text conversion using android platform," *International Journal of Engineering Research and Applications (IJERA)*, vol. 3, no. 1, pp. 253–258, 2013.

[9] G. Farias, S. Vergara, E. Fabregas, G. Hermosilla, S. Dormido-Canto, and S. Dormido, "Clasificador de noticias usando autoencoders," in *2018 IEEE International Conference on Automation/XXIII Congress of the Chilean Association of Automatic Control (ICA-ACCA)*, 2018, pp. 1–6.

[10] V. Jakkula, "Tutorial on support vector machine (svm)," *School of EECS, Washington State University*, vol. 37, no. 2.5, p. 3, 2006.

[11] T. Kumar, M. Mahrishi, and S. Nawaz, "A review of speech sentiment analysis using machine learning," in *Proceedings of Trends in Electronics and Health Informatics. Lecture Notes in Networks and Systems*, vol. 376, 2022, pp. 21–28.

[12] B. F., P. M., R. W.F., S. B., and W. B., "A database of german emotional speech," in *Proceedings Interspeech*, 2005.

[13] E. Handschin, "Electrical network control," *Control Systems, Robotics, and Automation*, vol. XVIII, 2005. [Online]. Available: https://www.eolss.net/sample-chapters/c18/E6-43-33-05.pdf

[14] S. Boyer, *SCADA: Supervisory Control and Data Acquisition*, 4th ed. International Society of Automation, 2010.

[15] H.-K. Podszeck, *Carrier Communication over Power Lines: Communication Problems in Electric System Operation*. Springer Berlin Heidelberg, 1972.

[16] E. Tsampasis, D. Bargiotas, C. Elias, and L. Sarakis, "Communication challenges in smart grid," *MATEC Web of Conferences*, vol. 41, p. 01004, 02 2016.

[17] J. Acevedo, G. Garcia, R. Ramirez, E. Fabregas, G. Hermosilla, S. Dormido-Canto, and G. Farias, "Uncertainty detection in supervisor–operator audio records of real electrical network operations," *Electronics*, vol. 13, no. 1, 2024. [Online]. Available: https://www.mdpi.com/2079-9292/13/1/141

[18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[19] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[20] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Computational statistics & data analysis*, vol. 52, no. 1, pp. 155–173, 2007.

[21] G. Robert, G. Seni, J. Elder, N. Agarwal, and H. Liu, *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictionsgoo*. Morgan & Claypool, 2010.

[22] The MathWorks Inc., "Matlab (r2019b)," Natick, Lakeside Campus, Massachusetts, United States.

[23] R. Benítez, G. Escudero, S. Kanaan, and D. Rodó, *Inteligencia artificial avanzada*. Editorial UOC, 2014.