risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 28/04/2025
Aceitação/Acceptance: 26/07/2025

Clasificación de Frases en Zapoteco mediante CNN con Espectrogramas

Mariano Martínez Patiño[1], Sergio Juárez Vázquez[1], Efraín Dueñas Reyes[1], Francisco Javier Sol Sampedro[1], Nicolas Hernández Ruiz[1]

mariano971149@gmail.com; sjuarez@sandunga.unistmo.edu.mx; eduenas@sandunga.unistmo.edu.mx; franciscosol@sandunga.unistmo.edu.mx; nicolas@sandunga.unistmo.edu.mx

[1] Universidad del Istmo, Ciudad Universitaria S/N, Santa Cruz, 70760 Santo Domingo Tehuantepec, Oax., México.

DOI: 10.17013/risti.59.36–52

Resumen: Este estudio se enfoca en la clasificación de espectrogramas, representaciones visuales del audio para aplicar aprendizaje automático. Los métodos tradicionales, como los MFCCs con clasificadores clásicos, presentan limitaciones en lenguas con pocos recursos, como el zapoteco del Istmo. Modelos avanzados como RNNs y transformers requieren grandes volúmenes de datos, difíciles de obtener en contextos indígenas. Como alternativa, se propone una red neuronal convolucional profunda de 28 capas, entrenada con 10 frases comunes convertidas en espectrogramas y aumentadas artificialmente. El modelo logró un 100% de precisión en entrenamiento y 96.2% en validación. Aunque prometedor, se destaca la necesidad de ampliar el conjunto de datos. El trabajo evidencia el potencial del aprendizaje profundo para mejorar la comunicación intercultural y preservar lenguas indígenas en peligro.

Palabras-clave: Comunicación intercultural; lenguas indígenas zapoteca; imágenes espectrales; red neuronal profunda.

Application of CNN in Audio Recognition of Isthmus Zapotec

Abstract: This study focuses on the classification of spectrograms, visual representations of audio, for the application of machine learning. Traditional methods, such as MFCCs with classical classifiers, have limitations in resource-poor languages such as Isthmus Zapotec. Advanced models, such as RNNs and transformers, require large volumes of data, which are often difficult to obtain in indigenous contexts. As an alternative, a 28-layer deep convolutional neural network is proposed, trained with 10 common phrases converted into spectrograms and artificially augmented. The model achieved 100% training accuracy and 96.2% validation accuracy. Although promising, the need to expand the dataset is highlighted. This work demonstrates the potential of deep learning to improve intercultural communication and preserve endangered indigenous languages.

Keywords: Intercultural communication; zapotec indigenous languages; spectral images; deep neural network.

1. Introducción

La lengua indígena zapoteco del Istmo, conocida localmente como Diidxazá, es una de las variantes más representativas de la familia lingüística zapoteca, hablada principalmente en la región del Istmo de Tehuantepec, Oaxaca, México. Esta lengua no solo cumple una función comunicativa, sino que se erige como pilar fundamental de la identidad, las tradiciones y la historia de los pueblos zapotecos.

Sin embargo, actualmente se encuentra en una situación crítica. De acuerdo con el Censo de Población y Vivienda 2020 del Instituto Nacional de Estadística y Geografía (INEGI, 2020), el número de hablantes de zapoteco del Istmo ha disminuido drásticamente en la última década, con reducciones que van del 1% y el 30% en municipios como Unión Hidalgo, Tehuantepec, Matías Romero, entre otros. Esta pérdida representa una amenaza para el patrimonio cultural y lingüístico de México, especialmente porque muchas personas mayores de estas comunidades, que no hablan español, se enfrentan a barreras de comunicación en contextos esenciales como los servicios médicos o administrativos (Delgadillo et al., 2020).

Esta dramática pérdida de hablantes representa una amenaza directa para el patrimonio cultural y lingüístico de México. La situación es particularmente crítica porque muchas personas mayores de estas comunidades, que no hablan español, se enfrentan a barreras de comunicación en contextos esenciales como los servicios médicos, administrativos y de atención social (Delgadillo et al., 2020). En particular, los adultos mayores monolingües en zapoteco enfrentan importantes barreras cuando intentan acceder a servicios básicos que requieren comunicación en español, lo que compromete significativamente su calidad de vida y acceso a atención adecuada. Por lo tanto, es crucial desarrollar herramientas de comunicación efectivas que puedan atender adecuadamente las necesidades de estos hablantes.

En respuesta a esta problemática, las tecnologías digitales se presentan como soluciones prometedoras para la preservación de las lenguas indígenas. Por ejemplo, plataformas de juegos serios y aplicaciones móviles con realidad aumentada han facilitado la enseñanza de lenguas ancestrales (Panamá-Mazhenda & Robles Bykbaev, 2024). De manera similar, el proyecto Etnoenglish Cultural Exchange promueve la revitalización de lenguas indígenas colombianas mediante contenidos digitales creados colaborativamente, integrando las TIC en la educación intercultural (Villa et al., 2024).

Entre 2010 y 2020, la brecha digital en México se ha reducido considerablemente, principalmente gracias a la adopción generalizada de teléfonos móviles, aunque las comunidades marginadas aún permanecen rezagadas. Paralelamente, investigaciones recientes han demostrado que las plataformas web pueden contribuir de manera efectiva a la revitalización de las lenguas indígenas (Medina et al., 2023).

Dentro de este panorama tecnológico, las redes neuronales convolucionales (CNN) son una solución eficaz para reconocer patrones complejos en señales acústicas mediante el análisis de espectrogramas, que son representaciones visuales en 2D de señales de audio (Franzoni, 2023). Un espectrograma muestra cómo varía la intensidad de las diferentes frecuencias de un sonido a lo largo del tiempo, convirtiendo una señal de audio en una imagen bidimensional que puede ser procesada por las CNN.

Las CNN aprovechan esta transformación aplicando filtros en capas convolucionales para identificar y extraer características relevantes, como patrones frecuenciales, texturas espectrales y formas específicas dentro de estos espectrogramas. Gracias a su estructura jerárquica y capacidad para aprender representaciones en distintos niveles de abstracción, las CNN son especialmente adecuadas para clasificar patrones vocales y lingüísticos en grabaciones de audio.

Por lo tanto, el presente trabajo explora el potencial de las CNN para la clasificación de frases en zapoteco del Istmo mediante la transformación de audios en espectrogramas, contribuyendo así a la preservación tecnológica de esta lengua indígena amenazada. Este enfoque permite convertir señales de audio en imágenes espectrales que pueden ser procesadas por la red neuronal, facilitando el reconocimiento automático de expresiones comunicativas. El modelo presentado tiene como objetivo principal realizar una clasificación automática de espectrogramas de audio correspondientes a frases comunes en zapoteco del Istmo, facilitando así la comunicación entre hablantes de esta lengua indígena y personas que no la dominan.

El documento está organizado de la siguiente forma: la Sección 2 expone los antecedentes y trabajos relacionados, proporcionando el contexto para el enfoque planteado. La Sección 3 describe el conjunto de datos utilizado y las muestras que contiene, incluyendo el proceso de transformación a espectrogramas. En la Sección 4 se explica la metodología, detallando la arquitectura de la red neuronal convolucional propuesta, las técnicas de preprocesamiento de datos y los parámetros de entrenamiento utilizados. En la Sección 5 se presentan los resultados experimentales, incluyendo métricas de precisión, matrices de confusión y análisis del rendimiento del modelo, seguidos de una interpretación de los hallazgos. Finalmente, la Sección 6 concluye el artículo, resumiendo las principales contribuciones y delineando perspectivas para futuras investigaciones.

2. Trabajos relacionados

Diversos estudios han demostrado que las redes neuronales convolucionales (CNN) son altamente eficaces para la clasificación de señales de audio, especialmente cuando estas se transforman en representaciones visuales como espectrogramas (Franzoni, 2023). El uso de espectrogramas permite que las CNN aprovechen su capacidad para detectar patrones espaciales, facilitando el reconocimiento de características acústicas relevantes en tareas como la clasificación de sonidos ambientales, música y habla.

Modelos basados en CNN han alcanzado precisiones superiores al 90% en conjuntos de datos estándar, y su rendimiento se potencia al combinarse con técnicas como Mel-Frequency Cepstral Coefficients (MFCC) o espectrogramas Mel (Mushtaq et al., 2021). Además, la transferencia de aprendizaje con arquitecturas preentrenadas en imágenes, como ResNet o DenseNet, ha demostrado ser efectiva para tareas de audio, incluso cuando los datos disponibles son limitados (Zaman et al., 2023).

Por otra parte, se han desarrollado modelos híbridos que combinan CNN con unidades recurrentes controladas (GRU), mejorando la captura de características espaciales y temporales en señales de audio. Estos modelos han sido aplicados con éxito en la identificación de hablantes (Ye & Yang, 2021). Asimismo, los sistemas de identificación

de idiomas han utilizado estas técnicas, alcanzando alta precisión incluso con segmentos cortos de audio (Rammo & Al-Hamdani, 2022).

En el ámbito de la clasificación de dialectos y lenguas regionales, recientes investigaciones han reportado resultados prometedores mediante enfoques de aprendizaje profundo. Por ejemplo, Al–Anzi & Thankaleela (2025) utilizaron un modelo preentrenado basado en redes neuronales convolucionales 1D y 2D para identificar dialectos árabes, clasificando clips de audio en dialectos como árabe egipcio, árabe del Golfo, árabe levantino, árabe moderno estándar y árabe saudí, entre otros, logrando una precisión de prueba del 94.28% y de validación del 95.55%.

Por su parte, Lai et al. (2024) emplearon coeficientes cepstrales en frecuencia Mel (MFCC) y una red neuronal convolucional 1D para extraer y clasificar características de género y dialecto regional en señales de habla, alcanzando precisiones superiores al 90% y altos valores en métricas de recall y F1 score, demostrando eficacia y robustez en reconocimiento automático.

En cuanto al reconocimiento de emociones en dialecto árabe saudí, Aljuhani et al. (2021) utilizaron MFCC y espectrogramas Mel para extraer características, aplicando clasificadores SVM, MLP y KNN, con SVM alcanzando la mejor precisión del 77.14%, mejorando el reconocimiento automático de emociones en árabe.

Sin embargo, una revisión sistemática destacó sesgos en la investigación, como la preferencia por variedades regionales más estudiadas y la escasez de recursos para dialectos urbanos y lenguas vernáculas (Elnagar et al., 2021). Estos vacíos evidencian la necesidad de ampliar los estudios hacia lenguas y dialectos menos explorados, como las lenguas indígenas.

En el contexto específico de lenguas indígenas y preservación lingüística, Salau et al. (2020) implementaron un modelo CNN-LSTM para la clasificación de acentos en tres lenguas indígenas nigerianas, utilizando características espectrales extraídas de grabaciones de audio. Su trabajo demuestra la viabilidad de aplicar técnicas de aprendizaje profundo a lenguas con recursos limitados, logrando una precisión del 78% en la clasificación de variantes dialectales. Esta investigación establece un precedente importante para el uso de CNN en el análisis de lenguas indígenas.

Asimismo, Noda et al. (2015) desarrollaron un sistema audio-visual basado en CNN aplicado a espectrogramas, logrando una precisión del 89% en tareas de reconocimiento de fonemas. Este enfoque subraya la riqueza informativa de las representaciones espectrales para la clasificación de patrones vocales, lo que es especialmente relevante para lenguas con características fonológicas específicas.

Finalmente, en el ámbito de la preservación de lenguas amenazadas mediante tecnología, Dueck (2024) presenta un marco comprehensivo para el uso de inteligencia artificial en la preservación de historias orales indígenas, enfatizando la importancia de sistemas automáticos de reconocimiento de habla adaptados a las particularidades fonológicas y estructurales de estas lenguas.

Los estudios revisados en esta sección demuestran la eficacia de las redes neuronales convolucionales aplicadas a espectrogramas en diversas aplicaciones de análisis de señales de audio, particularmente en tareas de clasificación y reconocimiento de habla.

Sin embargo, se destaca la limitada investigación específica sobre la aplicación de CNN a espectrogramas para la clasificación de frases completas en lenguas indígenas mesoamericanas como el zapoteco del Istmo, lo que sugiere un área de investigación valiosa y poco explorada que puede contribuir significativamente a la preservación de estas lenguas en peligro de extinción.

Por tanto, el presente trabajo busca abordar esta brecha, proponiendo un modelo basado en CNN para la clasificación de frases en zapoteco mediante el análisis de espectrogramas, aprovechando las ventajas del aprendizaje profundo en contextos con recursos limitados.

3. Descripción del conjunto de datos

En este estudio, la calidad y la cantidad de los datos son fundamentales para el éxito de la clasificación de frases en zapoteco mediante redes neuronales convolucionales (CNN). Por lo tanto, es crucial describir detalladamente cómo se recolectaron, preprocesaron y aumentaron los datos utilizados en este trabajo. A continuación, se presenta una explicación paso a paso de este proceso.

El conjunto de datos está formado por 5,055 grabaciones de 10 frases comunes en zapoteco del Istmo, seleccionadas para representar situaciones cotidianas de solicitud de ayuda y necesidades básicas de comunicación. Cada frase cuenta con entre 465 y 570 muestras. Esta base es fundamental para entrenar y evaluar la red neuronal convolucional (CNN). La distribución detallada se presenta en la Tabla 1.

Frase en español	Frase en zapoteco	Número de muestras
Tengo hambre	*Candaana'*	480
Tengo frio	*Cayaca' nanda'*	525
Tengo sed	*Cayate' nisa*	570
Tengo sueño	*Cadi candaana dia'*	495
No tengo Hambre	*Cadi Cayaca nanda dia'*	465
No tengo frio	*Cadi Cayate dia' nisa*	525
No tengo sed	*Hracaladxe' gahua'*	525
Quiero comer	*Hracaladxe' nisa*	495
Quiero agua	*Hracaladxe' gaaze'*	495
Quiero bañarme	*Dxa' bacaanda lua'*	480

Tabla 1 – Lista de frases utilizadas en el conjunto de datos.

Para organizar los datos, se creó una carpeta por cada frase (o clase), asignando un nombre que identifica la frase correspondiente, como se muestra en la Figura 1. Este paso aseguró una estructura clara para el manejo de los datos durante el entrenamiento del modelo.

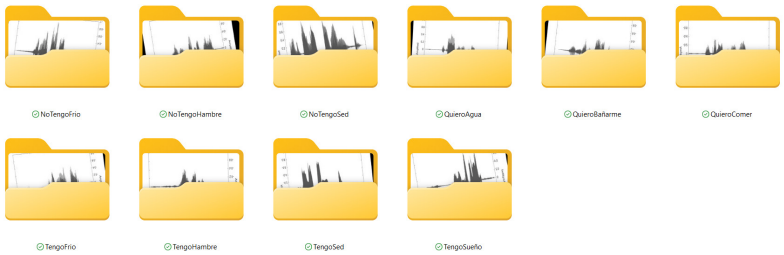

Figura 1 – Carpetas por clase del conjunto de datos.

Para que las CNN procesen las grabaciones de audio, estas se transformaron en espectrogramas utilizando MATLAB. Un espectrograma es una representación visual que muestra cómo varían las frecuencias de un audio a lo largo del tiempo, con las amplitudes indicadas mediante colores o intensidades. Estas imágenes capturan características acústicas clave, como tonos y patrones fonéticos, esenciales para distinguir frases en una lengua tonal como el zapoteco.

Posteriormente, los espectrogramas se convirtieron a escala de grises y se estandarizaron a un tamaño uniforme de 500 × 500 píxeles para simplificar el procesamiento y garantizar la consistencia en las entradas del modelo, lo cual es esencial para el entrenamiento efectivo de la CNN. Un ejemplo de espectrograma generado se presenta en la Figura 2.

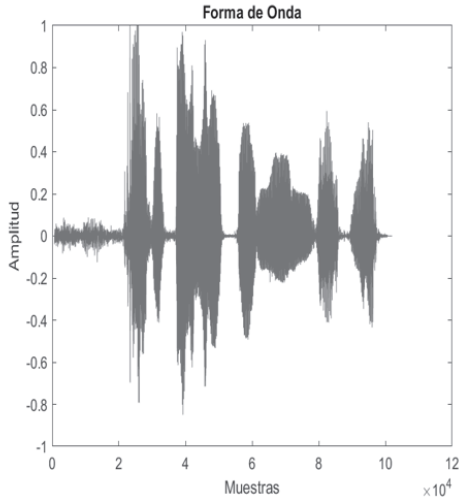

Figura 2 – Espectrograma de frase en zapoteco del Istmo

Dado que el número inicial de grabaciones por frase era limitado, se aplicó un aumento de datos para enriquecer el conjunto y mejorar la robustez del modelo. Utilizando un

software especializado para análisis espectral, se realizaron transformaciones en los espectrogramas, como rotaciones y traslaciones, para generar nuevas versiones de las imágenes sin alterar las características acústicas esenciales. Por ejemplo, una rotación de la imagen no cambia el contenido fonético del audio, pero introduce variabilidad que ayuda al modelo a generalizar mejor. La Tabla 2 detalla la cantidad de imágenes generadas por clase tras este proceso, mostrando un conjunto de datos más amplio y diverso.

Frase	Clase	Número de muestras
Tengo hambre	TengoHambre	960
Tengo frio	TengoFrio	1050
Tengo sed	TengoSed	1140
Tengo sueño	TengoSueño	990
No tengo Hambre	NoTengoHambre	930
No tengo frio	NoTengoFrio	1050
No tengo sed	NoTengoSed	1050
Quiero comer	QuieroComer	990
Quiero agua	QuieroAgua	990
Quiero bañarme	QuieroBañarme	960

Tabla 2 – Información del conjunto de datos.

Este proceso permitió crear un conjunto de datos robusto y estructurado, partiendo de un total inicial de 5,550 grabaciones. Tras aplicar aumento de datos mediante transformaciones en los espectrogramas, como rotaciones y traslaciones, se duplicó aproximadamente el número de muestras por clase, alcanzando un total de 10,110 grabaciones. Esta ampliación enriqueció la diversidad del conjunto, optimizando el rendimiento de la red neuronal y fortaleciendo la preservación y comunicación en esta lengua indígena.

4. Metodología

En este estudio se exploraron diversas arquitecturas de redes neuronales convolucionales (CNN) para optimizar la precisión en la clasificación de audios de frases en zapoteco del Istmo, evaluando variaciones en el número de capas, estructuras simétricas y asimétricas, y estrategias de reducción dimensional. Tras un análisis comparativo, se seleccionó una arquitectura eficiente de 28 capas, que inicia con una capa de entrada de imágenes de 500 por 500 píxeles con un canal.

La red incluye seis bloques convolucionales con sus respectivas capas de normalización por lotes, funciones de activación ReLU y capas de max pooling que reducen progresivamente la dimensión espacial y aumentan la profundidad de los mapas de características. Estas capas convolucionales tienen tamaños que van desde 500 x 500 x 8 hasta 10 x 10 x 256.

Finalmente, la arquitectura incluye dos capas totalmente conectadas y una capa softmax para la clasificación, con salida de tamaño 1 por 1 por 2, correspondiente a las clases definidas. Esta red profunda está diseñada específicamente para maximizar el rendimiento mediante un proceso efectivo de extracción de características y clasificación de frases en zapoteco. La Figura 3 ilustra el diagrama general de esta arquitectura.

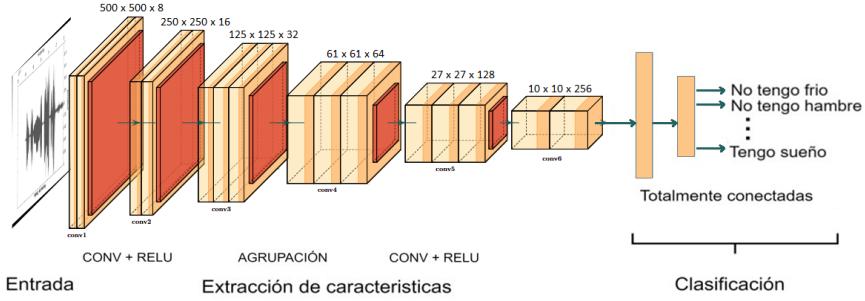

Figura 3 – Arquitectura de CNN para clasificación de espectrogramas

Las capas convolucionales son fundamentales para extraer características relevantes de los espectrogramas de audio. La arquitectura propuesta incluye seis capas convolucionales con un número creciente de filtros y tamaños de núcleo, lo que permite detectar patrones cada vez más complejos. La primera capa procesa espectrogramas en escala de grises de 500×500 píxeles utilizando ocho filtros de tamaño 3×3. Las siguientes capas aplican filtros de 16, 32, 64, 128 y 256, incrementando el tamaño de los núcleos hasta 6×6 en las capas más profundas. Esta progresión mejora la capacidad de la red para captar características acústicas abstractas. La figura 4 muestra ocho kernels representativos de cada una de estas capas convolucionales.

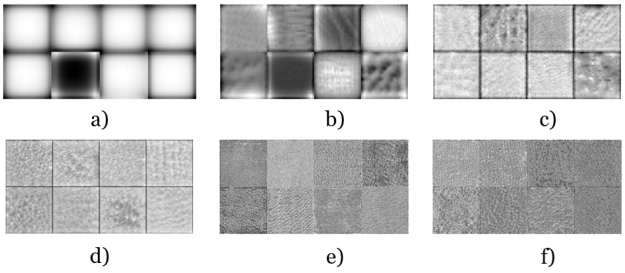

Figura 4 – Extracción de características en las capas: a) convolución 1, b) convolución 2, c) convolución 3, d) convolución 4, e) convolución 5 y f) convolución 6.

Las capas de activación ReLU introducen no linealidad al eliminar valores negativos y conservar los positivos, ayudando a la red a aprender las variaciones tonales propias del

zapoteco. El modelo incluye seis capas ReLU. La figura 5 ilustra algunos de los kernels asociados a estas capas ReLU.

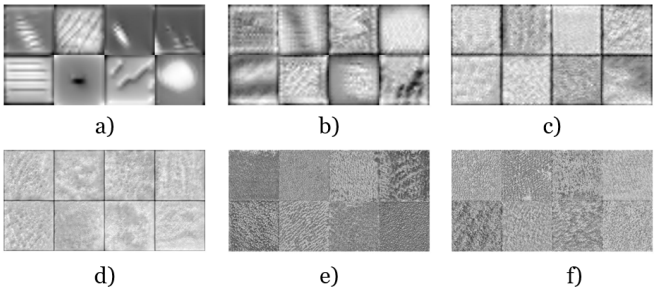

Figura 5 – Extracción de características en las capas: a) ReLU 1, b) ReLU 2, c) ReLU 3, d) ReLU 4, e) ReLU 5 y f) ReLU 6.

Por otro lado, las operaciones de max pooling, aplicadas en cinco etapas, reducen la dimensionalidad al seleccionar el valor máximo en regiones de 2×2, preservando las características esenciales y evitando el sobreajuste. La figura 6 presenta ocho kernels de algunas de estas capas de max pooling.

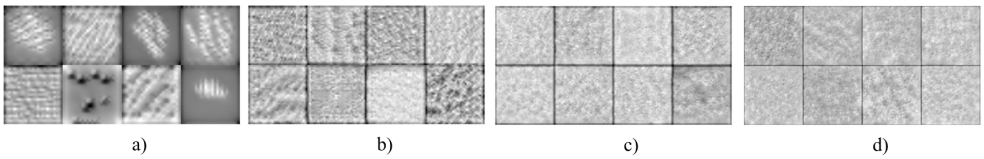

Figura 6 – Extracción de características en las capas: a) MaxPooling 1, b) MaxPooling 2, c) MaxPooling 3, d) MaxPooling 4.

5. Resultados

Esta sección presenta los resultados de la evaluación del modelo de red neuronal convolucional (CNN) entrenado para clasificar frases en zapoteco utilizando espectrogramas. El análisis incluye la distribución de los datos, los parámetros de entrenamiento, la precisión del modelo y las métricas de rendimiento basadas en matrices de confusión. El objetivo es evaluar la capacidad del modelo para generalizar y clasificar con precisión espectrogramas no vistos correspondientes a frases habladas en zapoteco.

Para el entrenamiento del modelo, el conjunto de datos se dividió en dos subconjuntos: un conjunto de entrenamiento que contiene el 80% de las imágenes y un conjunto de validación que contiene el 20% restante.

Clase	Entrenamiento	Validación
NoTengoFrio	840	210
NoTengoHambre	744	186
NoTengoSed	840	210
QuieroAgua	792	198
QuieroBañarme	768	192
QuieroComer	792	198
TengoFrio	840	210
TengoHambre	768	192
TengoSed	912	228
TengoSueño	792	198
Total	**8088**	**2022**

Tabla 4 – Número de muestras en conjunto de validación y entrenamiento.

Los parámetros utilizados para el proceso de entrenamiento fueron los siguientes: se empleó el algoritmo de gradiente descendente estocástico, un tamaño de lote de 40 muestras, un máximo de 20 épocas, una tasa de aprendizaje inicial de 0.0001 y una frecuencia de validación de 6 iteraciones. El entrenamiento se llevó a cabo durante 20 épocas, con 202 iteraciones en cada una, sumando un total de 4040 iteraciones en aproximadamente 627 minutos, utilizando un conjunto de 8088 imágenes. Estas pruebas se realizaron en una computadora equipada con un procesador Intel i5, GPU NVIDIA GeForce RTX 3050 y 32 GB de memoria RAM. El resultado obtenido fue un porcentaje de precisión del 96.24%, como se muestra en la figura 7.

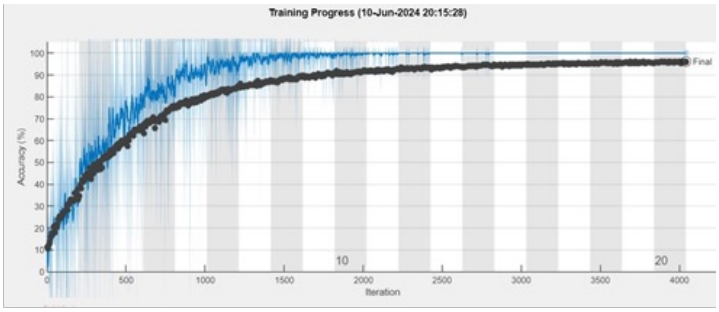

Figura 7 – Curva de aprendizaje resultante del proceso de entrenamiento.

Para evaluar el desempeño del modelo se usaron matrices de confusión en los conjuntos de entrenamiento y validación. Se analizaron tres métricas clave: Verdaderos Positivos (VP), que indican frases correctamente clasificadas; Falsos Positivos (FP), que son frases de otras clases erróneamente asignadas a una clase; y Falsos Negativos (FN),

que corresponden a frases mal clasificadas fuera de su clase original. Estas métricas permiten medir con detalle la precisión del modelo en la clasificación de las 10 clases de frases en zapoteco del Istmo.

En el conjunto de entrenamiento, el modelo alcanzó una precisión del 100%, clasificando correctamente todas las 8088 imágenes de espectrogramas correspondientes a las 10 frases en zapoteco del Istmo. Como se muestra en la Figura 8, la matriz de confusión presenta valores altos en la diagonal principal, indicando que no hubo errores de clasificación. Este rendimiento perfecto sugiere que la red neuronal convolucional (CNN) aprendió eficazmente los patrones acústicos de los espectrogramas. Sin embargo, una precisión del 100% en el conjunto de entrenamiento puede indicar sobreajuste, lo que resalta la importancia de evaluar el modelo con datos no vistos en el conjunto de validación para garantizar su capacidad de generalización.

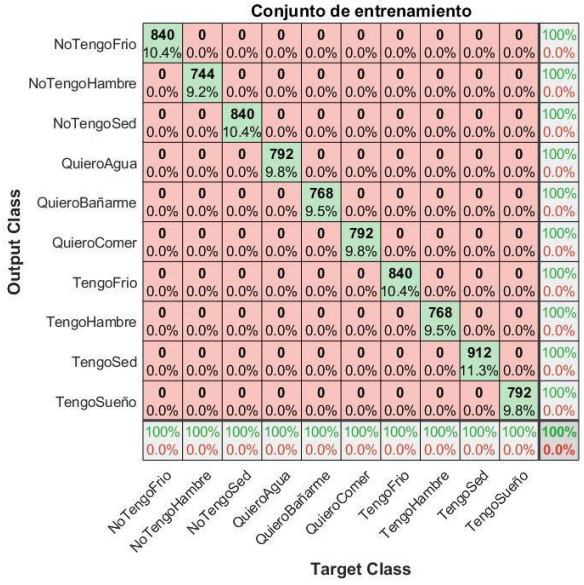

Figura 8 – Matriz de confusión del conjunto de entrenamiento.

Por otra parte, en el conjunto de validación, un total de 199 imágenes de la clase NoTengoFrio fueron clasificadas correctamente dentro de la misma clase, representando casos verdaderos positivos. Mientras tanto, 11 imágenes de otras clases fueron clasificadas incorrectamente como NoTengoFrio, correspondiendo a falsos positivos, y 5 imágenes que pertenecían a esta clase fueron clasificadas erróneamente como otras, representando falsos negativos. Del mismo modo, para la clase NoTengoHambre, el modelo identificó 180 verdaderos positivos, junto con 6 falsos positivos y 5 falsos negativos. La matriz de

confusión que resume estos resultados se muestra en la Figura 9, donde se observa una exactitud del 96,2% en el conjunto de validación.

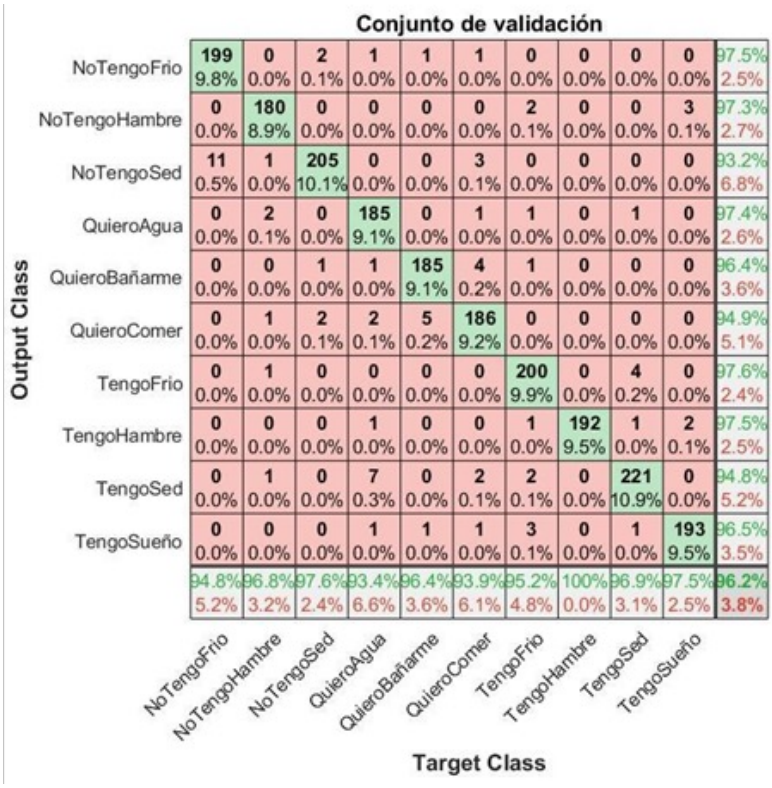

Figura 9 – Matriz de confusión del conjunto de validación.

En total, el conjunto de datos produjo un total de 1946 verdaderos positivos, 76 falsos positivos y 76 falsos negativos, como se detalla en la Tabla 5.

Clase	VP	FP	FN
NoTengoFrio	199	11	5
NoTengoHambre	180	6	5
NoTengoSed	205	5	15
QuieroAgua	185	13	5
QuieroBañarme	185	7	7
QuieroComer	186	12	10
TengoFrio	200	10	5

Clase	VP	FP	FN
TengoHambre	192	0	5
TengoSed	221	7	12
TengoSueño	193	5	7
Total	**1946**	**76**	**76**

Tabla 5 – Resultados de la matriz de confusión del conjunto de validación.

Además, se calcularon la precisión, la sensibilidad y la especificidad del modelo para proporcionar una evaluación más completa de su rendimiento. La precisión mide la proporción de predicciones correctas, calculada como (TP + TN) / (TP + TN + FP + FN). La sensibilidad (o recuerdo) refleja la capacidad del modelo para identificar correctamente los casos positivos y se calcula como TP / (TP + FN). La especificidad indica la capacidad del modelo para identificar correctamente los casos negativos y se calcula como TN / (TN + FP). Estas métricas se resumen en la Tabla 6.

Conjunto	Exactitud	Sensibilidad	Especificidad
Entrenamiento *99.32%* *99.29%* *99.35%*	100%	100%	100%
Validación	96.21%	96.23%	96.23%

Tabla 6 – Resultados Exactitud, Sensibilidad y Especificidad

6. Conclusiones

La preservación y facilitación de la comunicación en lenguas indígenas, como el zapoteco del Istmo, es fundamental para mejorar la integración y calidad de vida de las comunidades indígenas. En este estudio, se propuso un modelo de red neuronal convolucional (CNN) de 28 capas para la clasificación de espectrogramas de 10 frases en zapoteco del Istmo. El modelo fue entrenado y validado con un conjunto de datos que contiene 10,110 imágenes (8088 de entrenamiento y 2022 de validación), generadas a partir de un número limitado de audios originales, normalizadas y ajustadas a un tamaño de 500 x 500 píxeles.

Los resultados demuestran que el modelo CNN propuesto clasifica con alta precisión frases en zapoteco a partir de espectrogramas, alcanzando un 100% de exactitud en entrenamiento y un 96.2% en validación. Aunque el conjunto de datos es pequeño y limitado a 10 frases, estos resultados son comparables e incluso superiores a los obtenidos en otros trabajos de clasificación de audio con aprendizaje profundo en contextos de recursos limitados.

De manera similar, Telmem et al. (2025) lograron un 85% de precisión en el reconocimiento del habla amazigh utilizando CNNs basadas en espectrogramas. Wang et al. (2020) alcanzaron un 92.1% en la clasificación de escenas acústicas combinando

múltiples representaciones espectrales. De forma similar, Binjaku et al. (2022) desarrollaron un sistema para identificar el idioma albanés en grabaciones de voz mediante espectrogramas y CNN, obteniendo una precisión del 94%.

Otros estudios relevantes incluyen a Dwivedi et al. (2023), quienes lograron más del 95% de precisión en la identificación de tartamudez en habla hindi usando modelos secuenciales de CNN entrenados con espectrogramas. Lesnichaia et al. (2022) alcanzaron entre 96.4% y 98.7% de precisión en la clasificación de acentos en inglés, empleando CNNs con mel-espectrogramas de amplitud en escala lineal, superando modelos previos con el mismo conjunto de datos. Por su parte, Dayal et al. (2022) lograron un 95.2% de precisión en la clasificación de 11 sonidos de fondo incrustados en señales de habla humana, utilizando un modelo CNN ligero basado en espectrogramas, que además superó a modelos de referencia en tiempo de inferencia.

Por su parte, Demir et al. (2020) desarrollaron dos modelos, uno basado en redes neuronales artificiales (ANN) y otro en CNN, para la identificación del idioma albanés a partir de espectrogramas, alcanzando precisiones del 85% y 94%, respectivamente.

Una limitación importante de este estudio es el reducido tamaño del conjunto de audios originales por clase, lo cual puede afectar la capacidad de generalización del modelo CNN en la clasificación de frases en zapoteco. La literatura especializada destaca que tanto el tamaño como la diversidad del conjunto de datos son factores clave para el rendimiento y la robustez de los modelos de aprendizaje profundo aplicados a la clasificación de audio.

Diversos estudios recientes han confirmado que aumentar el tamaño del dataset mejora la precisión y la generalización de los modelos CNN, aunque este beneficio tiende a estabilizarse tras cierto umbral (Paul et al., 2025; Shoumy et al., 2021). Además, no solo la cantidad sino también la diversidad de los datos —incluyendo diferentes hablantes, entonaciones y contextos acústicos— resulta fundamental para que el modelo sea efectivo en aplicaciones prácticas, como sistemas de traducción o asistencia en comunidades monolingües (Paul et al., 2025; Pandian et al., 2024).

Esta limitación puede reducir la capacidad del modelo para distinguir entre las variantes del zapoteco, afectando su precisión y utilidad en aplicaciones prácticas. Por ello, es necesario contar con un conjunto de datos amplio y representativo que incluya la diversidad lingüística y acústica de estas variantes para mejorar la generalización del modelo.

En futuras investigaciones, se recomienda explorar técnicas modernas como redes neuronales recurrentes o transformadores para mejorar la precisión y capacidad de generalización de nuestro modelo multiclase. Además, contar con un conjunto de datos más amplio y diverso que refleje la variedad lingüística y acústica del zapoteco fortalecerá su aplicabilidad en contextos reales.

Referencias

Al–Anzi, F. S., & Thankaleela, B. S. S. (2025). Region-Wise Recognition and Classification of Arabic Dialects and Vocabulary: A Deep Learning Approach. *Applied Sciences*, *15(12)*, 6516. https://doi.org/10.3390/app15126516

Aljuhani, R. H., Alshutayri, A., & Alahdal, S. (2021). Arabic speech emotion recognition from Saudi dialect corpus. *IEEE Access*, *9*, 127081-127085. https://doi.org/10.1109/ACCESS.2021.3110992

Binjaku, K., Janku, J., & Meçe, E. K. (2022). Identifying Low-Resource Languages in Speech Recordings through Deep Learning. *In 2022 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)* (pp. 1-6). IEEE. https://doi.org/10.23919/SoftCOM55329.2022.9911376

Dayal, A., Yeduri, S. R., Koduru, B. H., Jaiswal, R. K., Soumya, J., Srinivas, M. B., & Cenkeramaddi, L. R. (2022). Lightweight deep convolutional neural network for background sound classification in speech signals. *The Journal of the Acoustical Society of America*, *151(4)*, 2773-2786. https://doi.org/10.1121/10.0010257

Delgadillo, L. G., Arce, J. & Pastrana, S. A. (2020). Vulnerabilidad de la lengua en hablantes indígenas, el caso de México. *Circula*, *(12)*, 19–40. https://doi.org/10.17118/11143/18441

Demir, F., Abdullah, D. A., & Sengur, A. (2020). A new deep CNN model for environmental sound classification. *IEEE Access*, *8*, 66529-66537. https://doi.org/10.1109/ACCESS.2020.2984903

Dueck, G. W. (2024). Using AI to help preserve indigenous oral histories. *Proceedings of the 2024 IEEE International Humanitarian Technologies Conference*, 1-6. https://doi.org/10.1109/IHTC61819.2024.10855026

Dwivedi, S., Ghosh, S., & Dwivedi, S. (2023). Binary classifier for identification of stammering instances in Hindi speech data. *International Journal of Speech Technology*, *26(3)*, 765-774. https://doi.org/10.1007/s10772-023-10046-9

Elnagar, A., Yagi, S. M., Nassif, A. B., Shahin, I., & Salloum, S. A. (2021). Systematic literature review of dialectal Arabic: identification and detection. *IEEE Access*, *9*, 31010-31042. https://doi.org/10.1109/ACCESS.2021.3059504

Franzoni, V. (2023). Cross-domain synergy: Leveraging image processing techniques for enhanced sound classification through spectrogram analysis using CNNs. *Journal of Autonomous Intelligence*, *6(3)*, 1-14. https://doi.org/10.32629/jai.v6i3.678

INEGI (2020). Censo de Población y Vivienda 2020. https://www.inegi.org.mx/programas/ccpv/2020/#documentacion

Lai, H. Y., Hu, C. C., Wen, C. H., Wu, J. X., Pai, N. S., Yeh, C. Y., & Lin, C. H. (2024). Mel-Scale Frequency Extraction and Classification of Dialect-Speech Signals with 1D CNN based Classifier for Gender and Region Recognition. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3430296

Lesnichaia, M., Mikhailava, V., Bogach, N., Lezhenin, Y., Blake, J., & Pyshkin, E. (2022). Classification of Accented English Using CNN Model Trained on Amplitude Mel-Spectrograms. *In Interspeech* (pp. 3669-3673). https://doi.org/10.21437/Interspeech.2022-462

Medina, M. A. G., Jiménez, J. L. M., Meza, D. D. J. A., Vergara, J. T., & Cantero, C. L. (2023). Resignificación de la lengua materna Zenú mediante la plataforma web Tozí. RISTI - *Revista Ibérica de Sistemas e Tecnologias de Informação*, *(E59)*, 24-38. https://dialnet.unirioja.es/servlet/articulo?codigo=10079679

Mushtaq, Z., Su, S. F., & Tran, Q. V. (2021). Spectral images based environmental sound classification using CNN with meaningful data augmentation. *Applied Acoustics*, *172*, 107581. https://doi.org/10.1016/j.apacoust.2020.107581

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. (2015). Audio-visual speech recognition using deep learning. *Applied Intelligence*, *42(4)*, 722-737. https://doi.org/10.1007/s10489-014-0629-7

Panamá-Mazhenda, K., & Robles-Bykbaev, V. (2024). Revisión sistemática de literatura de metodologías para el diseño y desarrollo de juegos serios: análisis MLR. *Revista Ibérica de Sistemas e Tecnologias de Informação*, *(E66)*, 515-527. https://dialnet.unirioja.es/servlet/articulo?codigo=10003357

Pandian, J. A., Thirunavukarasu, R., & Kotei, E. (2024). A novel convolutional neural network model for automatic speaker identification from speech signals. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3385858

Paul, B., Phadikar, S., Bera, S., Dey, T., & Nandi, U. (2025). Isolated word recognition based on a hyper-tuned cross-validated cnn-bilstm from mel frequency cepstral coefficients. *Multimedia Tools and Applications*, *84(17)*, 17309-17328. https://doi.org/10.1007/s11042-024-19750-3

Rammo, F. M., & Al-Hamdani, M. N. (2022). Detecting the speaker language using CNN deep learning algorithm. *Iraqi Journal for Computer Science and Mathematics*, *3(1)*, 43-52. https://doi.org/10.52866/ijcsm.2022.01.01.005

Salau, A. O., Olowoyo, T. D., & Akinola, S. O. (2020). Accent classification of the three major Nigerian indigenous languages using 1D CNN LSTM network model. *In Advances in Computational Intelligence and Robotics (pp. 1-15). Springer*. https://doi.org/10.1007/978-981-15-2620-6_1

Shoumy, N. J., Ang, L. M., Rahaman, D. M. M., Zia, T., Seng, K. P., & Khatun, S. (2021). Augmented Audio Data in Improving Speech Emotion Classification Tasks. *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, *12799 LNAI*, 360–365. https://doi.org/10.1007/978-3-030-79463-7_30

Telmem, M., Laaidi, N., & Satori, H. (2025). The impact of MFCC, spectrogram, and Mel-Spectrogram on deep learning models for Amazigh speech recognition system. *International Journal of Speech Technology*, 1-14. https://doi.org/10.1007/s10772-025-10183-3

Villa, M. G. R., Zapata, J. A. S., Ospina-Giraldo, M. N., Holguin, M. M. O., Cataño, D. F. G., & Buitrago, J. D. R. (2024). Proyecto Etnoenglish Cultural Exchange: Una Experiencia Pedagógica de Digiculturalidad y Educación Inclusiva en la escuela. RISTI - *Revista Ibérica de Sistemas e Tecnologias de Informação*, *(E72)*, 370-381. https://dialnet.unirioja.es/servlet/articulo?codigo=9852073

Wang, H., Zou, Y., & Chong, D. (2020). Acoustic scene classification with spectrogram processing strategies. *arXiv preprint arXiv:2007.03781*. https://doi.org/10.48550/arXiv.2007.03781

Ye, F., & Yang, J. (2021). A deep neural network model for speaker identification. *Applied Sciences*, *11(8)*, 3603. https://doi.org/10.3390/APP11083603

Zaman, K., Sah, M., Direkoglu, C., & Unoki, M. (2023). A survey of audio classification using deep learning. *IEEE Access*, *11*, 106620-106649. https://doi.org/10.1109/ACCESS.2023.3318015