EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

INSTITUTO TECNOLÓGICO DE CD. GUZMÁN

TESIS
MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

TEMA DE TESIS:
RECONOCIMIENTO DE FENÓMENOS PARALINGÜÍSTICOS EN NIÑOS PARA SISTEMAS DE INTERACCIÓN BASADOS EN VOZ

QUE PARA OBTENER EL TITULO DE:
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:
ISABEL GUADALUPE VÁZQUEZ GÓMEZ

DIRECTORES:
DR. DANIEL FAJARDO DELGADO
DR. HUMBERTO PÉREZ ESPINOSA

CD. GUZMÁN JALISCO, MÉXICO, NOVIEMBRE DE 2021

Ciudad Guzmán, 25/noviembre/2021
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN
Asunto: *Autorización de impresión de Tesis*

ING. ISABEL GUADALUPE VÁZQUEZ GÓMEZ
CANDIDATA AL GRADO DE MAESTRO EN CIENCIAS DE LA COMPUTACIÓN
PRESENTE

De acuerdo con los Lineamientos para la Operación de los Estudios de Posgrado en el Tecnológico Nacional de México y las disposiciones en este Instituto, habiendo cumplido con todas las indicaciones que la Comisión Revisora realizó con respecto a su Trabajo de Tesis titulado "Reconocimiento de fenómenos paralingüísticos en niños para sistemas de interacción basados en voz", la División de Estudios de Posgrado e Investigación de este Instituto, concede la Autorización para que proceda a la impresión del mismo.

Sin otro particular, quedo de Usted.

A T E N T A M E N T E
Excelencia en Educación Tecnológica
"Innova, Transforma y Crea para ser Grande"

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
INSTITUTO TECNOLÓGICO DE CD. GUZMÁN
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

DRA. MARÍA GUADALUPE SÁNCHEZ CERVANTES
JEFA DE LA DIVISION DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

ccp. Archivo
MGSS/megg

Dedicatoria

A mis padres

Por ser el pilar fundamental para forjarme en lo profesional. Me han dado todo lo que soy como persona, mis valores, mis principios, mi perseverancia y mi empeño.

A mi pareja

Por creer en mí e impulsarme a continuar con mis sueños, por su paciencia y el cariño que me ha brindado de principio a fin.

Agradecimientos

Al Dr. Daniel Fajardo Delgado, por su tiempo y orientación durante mi trabajo, por ser un buen maestro y amigo.

Al Dr. Humberto Pérez Espinosa por orientarme y compartir sus conocimientos, apoyo y tiempo.

Gracias a mis compañeros y amigos del ITCG por los recuerdos alegres dentro de esta gran etapa profesional.

Al CONACyT por el apoyo y la oportunidad de realizar mis estudios de maestría.

Contenido

Página

Contenido (continuación)

Lista de figuras

Lista de tablas

Lista de tablas (continuación)

Lista de tablas (continuación)

Lista de tablas (continuación)

Tabla Página

Lista de pseudo-códigos

Pseudo-código Página

Capítulo I

Introducción

La paralingüística es el estudio de los aspectos no verbales del habla, los cuales comunican o matizan el sentido de lo que se dice. Estos aspectos pueden ser cualidades físicas del sonido tales como el tono, el timbre y la intensidad con la que se dicen las cosas. También pueden incluir vocalizaciones de escaso contenido léxico pero con un gran valor funcional, como el ¡Ah!, ¡Uy! y ¡Ajá!, entre otros. Otros aspectos incluyen reacciones fisiológicas o emocionales como la risa, el suspiro, el bostezo, el llanto, etc. Recientemente, todos estos aspectos o fenómenos paralingüísticos están tomando cada vez más relevancia en la interacción humano-máquina, así como en las áreas de procesamiento de señales y de aprendizaje máquina (Schuller *et al.*, 2013a; Vinciarelli *et al.*, 2008).

La computación paralingüística es un área de estudio emergente en la cual se investigan métodos computacionales que permitan obtener información de rasgos y atributos de las personas a partir del análisis acústico de sus voces y de la forma en que se expresan. El reconocimiento automático de fenómenos paralingüísticos puede ser útil para definir un perfil de usuario que permita adaptar y personalizar las interfaces de un sistema de interacción basado en voz. También puede ser útil para evaluar la calidad de la interacción basada en el habla entre el usuario y el sistema. Usando esta información, se puede contextualizar y adecuar la respuesta de un sistema de este tipo.

Existe una gran variedad de aplicaciones de computación paralingüística para el reconocimiento de emociones (Schuller *et al.*, 2009, 2013b, 2016), rasgos de usuarios

(Schuller *et al.*, 2012), autoctonía (Schuller *et al.*, 2015), entre otros. Sin embargo, la mayoría de ellas se centran en adultos y muy pocas están dirigidas a niños (Yildirim *et al.*, 2011; Pérez-Espinosa *et al.*, 2018b). Esto quizá se debe en parte a la escasez de corpus de datos adecuados para el estudio de fenómenos paralingüísticos del habla infantil. Recientemente, Pérez-Espinosa *et al.* (2018b) proponen un corpus de datos con grabaciones de audio de niños en distintos rangos de edades como una plataforma de experimentación para construir modelos de clasificación de fenómenos paralingüísticos.

El corpus de datos que se propone en (Pérez-Espinosa *et al.*, 2018b) contiene un total de $30,631$ audios de niños con información tanto objetiva como subjetiva de ellos. La información objetiva consta de la edad y el género de los niños, mientras que la información subjetiva trata de las emociones, actitudes, estado mental y aspectos paralingüísticos que se reflejan en cada audio. Esta información se define a través de diversas etiquetas asignadas por un conjunto de especialistas que, bajo su percepción e interpretación, describen características y atributos del estado mental, emocional y actitudinal de cada niño. Si bien se reportan casos donde los especialistas logran un consenso en el etiquetado de estos atributos subjetivos, en la mayoría de los audios esto no es así. En cambio se puede contar el número de coincidencias en el etiquetado para otorgar distintos niveles de confiabilidad de los mismos. Así, se pueden desarrollar métodos que ayuden a generar modelos de clasificación considerando los diferentes niveles de confiabilidad del etiquetado.

La presente tesis trata del desarrollo de un método de reconocimiento de fenómenos paralingüísticos en niños utilizando el corpus de datos de Pérez-Espinosa *et al.* (2018b). Debido a los diferentes niveles de confiabilidad en las etiquetas, se plantea el uso de técnicas semi-supervisadas de aprendizaje máquina. Con estas técnicas se buscó mejorar la confiabilidad del etiquetado para generar mejores modelos de clasificación de este tipo

de fenómenos.

I.1. Planteamiento del problema

Aún cuando en años recientes se ha logrado un avance en tareas relacionadas con el reconocimiento automático de fenómenos paralingüísticos, sigue siendo un reto muy importante la cantidad y naturaleza de los datos disponibles utilizados en la construcción de modelos de reconocimiento para dichos fenómenos. Además, el etiquetado de datos con fenómenos paralingüísticos no es trivial, ya que depende de la interpretación personal del etiquetador y generalmente presenta aspectos subjetivos.

Se han desarrollado diversos enfoques para abordar la subjetividad de los datos. Recientemente, Rizos y Schuller (2020) hacen referencia a la colaboración abierta distribuida (o *crowdsourcing*) como un medio para el etiquetado múltiple, permitiendo así ofrecer una solución rápida y minimizar el costo de dicho proceso. Por otra parte, Zhang *et al.* (2019) proponen el uso del aprendizaje cooperativo dinámico (DCL) en donde las instancias predichas con alta confianza se etiquetan automáticamente por la computadora, mientras que las instancias con mediana confianza están sujetas a revisión humana. El corpus de datos de Pérez-Espinosa *et al.* (2018b) minimiza la subjetividad mediante el etiquetado múltiple de la misma instancia por varios colaboradores. Este tipo de etiquetado múltiple hace aún más costoso el proceso y, por ende, este tipo de corpus de datos son escasos.

Por otro lado, gran parte del esfuerzo en el reconocimiento automático de información paralingüística se enfoca en adultos, y casi no se han desarrollado estudios para la clasificación de información paralingüística en niños (Yildirim *et al.*, 2011). Eso a pesar que diversas aplicaciones podrían utilizar esta información para el estudio del

desarrollo del niño, su educación, o asistencia en terapias de autismo, por nombrar algunas. Sólo un pequeño grupo de iniciativas de investigación ha abordado las problemáticas y retos asociados al estudio de la paralingüística computacional en niños, como son la generación de recursos y herramientas, la subjetividad en el etiquetado de los datos, la identificación de características acústicas apropiadas, y el desarrollo de clasificadores. En la medida que se aborden y solventen estos retos será posible crear sistemas y aplicaciones robustos que atiendan las necesidades de la población infantil.

I.2. Objetivos

A continuación se presentan los objetivos de este trabajo de investigación.

I.2.1. Objetivo general

El objetivo general del presente trabajo es desarrollar un método de clasificación de fenómenos paralingüísticos en niños que pueda emplearse para definir perfiles de usuarios en sistemas de interacción basadas en voz.

I.2.2. Objetivos específicos

Los objetivos específicos que coadyuvaron al cumplimiento del objetivo general son los siguientes:

- Disminuir el impacto negativo de la subjetividad en el etiquetado de fenómenos paralingüísticos mediante la aplicación de técnicas de aprendizaje máquina con un enfoque semi-supervisado.

- Construir una plataforma de experimentación y de construcción de modelos de reconocimientos de fenómenos paralingüísticos en niños utilizando datos etiquetados con niveles altos y bajos de confiabilidad, como los del corpus de Pérez-Espinosa *et al.* (2018b).
- Identificar las características acústicas más apropiadas para la clasificación de fenómenos paralingüísticos en niños.
- Implementar una aplicación que genere un perfil del niño basado en información paralingüística en el contexto de un sistema de interacción basada en voz.
- Desarrollar un método basado en aprendizaje semi-supervisado que permita obtener modelos más precisos que los métodos obtenidos con modelos de aprendizaje máquina supervisados.

I.3. Preguntas de investigación

Con base a los objetivos planteados, se formularon la siguientes preguntas de investigación:

- ¿Qué método de aprendizaje semi-supervisado es el más adecuado para generar un modelo de reconocimiento de fenómenos paralingüísticos en niños?
- ¿Es posible mejorar la precisión de los modelos de clasificación usando datos etiquetados con niveles bajos de consenso?
- ¿Qué proporción de datos etiquetados con alto nivel de consenso y con bajo nivel de consenso dan los mejores resultados?

- ¿Qué características acústicas son las más adecuadas para cada fenómeno paralingüístico estudiado?

I.4. Aportaciones

Las principales contribuciones de esta tesis son las siguientes:

- Un modelo de clasificación de fenómenos paralingüísticos en niños a través de técnicas de aprendizaje máquina semi-supervisado.
- Un prototipo para definir el perfil del niño en un sistema de interacción basada en voz.
- Nuevo conocimiento en el área de computación paralingüística para niños hispanohablantes.

I.5. Organización de la tesis

El resto del presente documento se organiza de la siguiente manera. El Capítulo II presenta el marco teórico de los fenómenos paralingüísticos, el aprendizaje máquina y aprendizaje semi-supervisado, introduciendo a las definiciones básicas y terminología que se utiliza a lo largo de este documento. El Capítulo III habla de la metodología que se llevo acabo para estructurar y equilibrar el corpus de datos, así como la extracción de características de los audios y la implementación del modelo dependiente e independiente. El Capítulo IV describe los resultados experimentales obtenidos, las características y las técnicas con las cuales se obtuvieron mejor resultado. El Capítulo V habla acerca del prototipo en el que se estuvo trabajando para hacer la implementación

de los modelos, y por último en el Capítulo VI se verán las conclusiones del presente trabajo, así como también el trabajo futuro a realizar.

Capítulo II

Marco teórico

A continuación se explican los conceptos y definiciones básicas que enmarcan el presente trabajo y que se derivan de tres áreas principales: los fenómenos paralingüísticos, el aprendizaje máquina y el aprendizaje semi-supervisado.

II.1. Fenómenos paralingüísticos

La paralingüística es el estudio de fenómenos acústicos (sonido y ondas sonoras) o lingüísticos (signos del lenguaje) modulados o embebidos dentro de un mensaje verbal (Abercrombie, 1968; Schuller *et al.*, 2013a). Además del paralenguaje, otros sistemas de comunicación no verbal reconocidos hasta el momento son la quinésica, la proxémica y la cronémica. La quinésica trabaja con el conjunto de gestos y movimientos corporales; la proxémica estudia la distancia corporal y la postura; y la cronémica es la duración de los signos con los que se comunican. De ellos, solamente el paralenguaje y la quinésica son considerados sistemas básicos o primarios por su implicación directa en cualquier acto de comunicación humana (Cestero Mancera, 2006). Cuando todos estos sistemas se consideran dentro de la comunicación, se conoce como procesamiento multi-modal (Schuller *et al.*, 2013a).

Los elementos o fenómenos paralingüísticos pueden expresarse a través de cualidades y modificadores fónicos, indicadores sonoros de reacciones fisiológicas y emocionales, elementos cuasi-léxicos, pausas y silencios que, en conjunto, dan significado a un mensaje hablado (Cestero Mancera, 2006; Rodríguez, 2007). Las cualidades y modificadores

fónicos (tono, timbre, intensidad y duración) son cualidades físicas o componentes acústicos del sonido que pueden determinar o precisar información (Rodríguez, 2007); e.g., un "sí, claro", dependiendo del tono con el que se emita, puede expresar acuerdo, desacuerdo, agrado, desilusión, etc. Las reacciones fisiológicas y emocionales son sonidos que comunican estados de ánimo en general, pero algunos tienen también la función de calificar enunciados o regular la conversación (Cestero Mancera, 2006); e.g., la risa, que ademas de denotar alegría, miedo o nerviosismo, también se utiliza para mostrar acuerdo o entendimiento. Respecto a los elementos cuasi-léxicos, se refiere a aquellos sonidos que, sin tener un nombre o una grafía establecidos, se utilizan convencionalmente con un valor comunicativo similar al de determinados signos lingüísticos (Poyatos, 1993). Estos pueden ser inactivos (como un "Chss" para llamar la atención), referenciales (como "Uuuh" para indicar lejanía) y expresivos (como "Aaahh" para mostrar bienestar físico o anímico).[1] Finalmente, las pausas y los silencios son pocos frecuentes en el español, pero pueden ser confirmadores de enunciados previos o para regular el cambio de turno en una conversación, entre otros. Para mayor referencia del paralenguaje ver (Poyatos, 1993; Quilis y Fernández, 1969). La paralingüística computacional es un área de estudio emergente en la cual se investigan métodos computacionales que permitan obtener información del estado, rasgos, y atributos de las personas a partir del análisis acústico del habla y otras vocalizaciones y tiene como objetivo reconocer significados importantes en la comunicación como actitudes, emociones e intenciones del hablante por medio de la información lingüística (verbal) para lograr una mejor interacción humano-máquina (Ishi *et al.*, 2008). Su área de estudio se interesa por

[1] De acuerdo a Cestero Mancera (2006), la diferencia entre pausa y silencio es el periodo de tiempo de ausencia de habla. Mientras que para la pausa se considera un periodo entre cero y un segundo, los silencios se consideran periodos mayores a un segundo.

proporcionar modelos computacionales de diferentes tipos de fenómenos lingüísticos que permiten obtener información del estado, rasgos y atributos de las personas a partir del análisis acústico de su voz y de la forma en que se expresan. Estos modelos se pueden basar en el conocimiento o en datos, aunque en algunos casos el trabajo lingüístico está motivado por el sentido científico que trata de dar explicación computacional a un fenómeno lingüístico en particular. De hecho, la realización de estos trabajos pueden ser incluidos en el reconocimiento de voz (Schuller y Batliner, 2013). En la comunicación humana existen dos elementos muy importantes el aspecto vocal, es decir los sonidos que emitimos con nuestro aparato fonador (e.g., voz, risas, tos, pausas llenas, gruñidos) y el aspecto verbal, es decir las palabras que usamos. En la figura 1 se muestran 4 categorías de fenómenos en la comunicación humana de acuerdo a la combinación de los factores vocales y verbales. De acuerdo a los trabajos seminales en esta área de estudio, la paralingüística computacional estudia las categorías I y II de esta figura, la categoría III la estudia parcialmente y la categoría IV queda totalmente fuera de su área de estudio.

Figura 1. Alcance de los fenómenos estudiados por la Paralingüística Computacional.

II.2. Aprendizaje máquina

En el contexto de aprendizaje máquina, y siguiendo la notación descrita en Zhu y Goldberg (2009a), una instancia $\vec{x}$ representa un objeto específico de decisión. La instancia es a menudo representada por un *vector de características* $\vec{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D$ de D dimensiones, donde cada dimensión se denomina *característica*. La longitud D del vector de características se conoce como la dimensionalidad del vector de características. Un *conjunto de entrenamiento* es una colección de instancias $\{\vec{x}\}_{i=1}^{n} = \{\vec{x}_1, \ldots, \vec{x}_n\}$ que representan la entrada del proceso de aprendizaje; i.e., la 'experiencia' que se le otorga al algoritmo de aprendizaje.

Los algoritmos de aprendizaje no supervisado trabajan sobre un conjunto de entrenamiento con n instancias $\{\vec{x}\}_{i=1}^{n}$ en el que no se provee de una supervisión de cómo las instancias individuales deben tratarse. Algunas tareas de aprendizaje no supervisado incluyen:

- agrupamiento (o clustering), donde el objetivo es separar las n instancias en grupos;
- detección de novedad, el cual identifica las instancias que son diferentes de la mayoría; y,
- reducción de dimensionalidad, que permite representar cada instancia con un vector de características de menor dimensionalidad mientras mantiene las características importantes del conjunto de entrenamiento.

Una *etiqueta* y es una predicción deseada de una instancia x. Las etiquetas pueden venir de un conjunto finito de valores llamadas *clases*, las cuales comúnmente se

codifican en números enteros. Para problema de más de dos clases, una codificación tradicional es $y \in \{1, \ldots, C\}$, donde C es el número de clases.

En el *aprendizaje supervisado*, el conjunto de entrenamiento consiste de pares, cada uno conteniendo una instancia $\vec{x}$ y una etiqueta $y : \{(\vec{x}_i, y_i)\}_{i=1}^{n}$. Estos pares (instancia, etiqueta) se denominan *datos etiquetados*. Sea X el dominio de instancias y Y el dominio de etiquetas, y dado un conjunto de entrenamiento $\{(\vec{x}_i, y_i)\}_{i=1}^{n}$, el aprendizaje supervisado entrena una función $f : X \to Y$ de alguna familia de funciones F con el objetivo de predecir la etiqueta verdadera y de datos futuros $\vec{x}$.

Como su nombre lo sugiere, el *aprendizaje semi-supervisado* es un intermedio entre el aprendizaje supervisado y el no supervisado. Aún así, existen dos enfoques de aprendizaje semi-supervisado ligeramente distintos: el aprendizaje semi-supervisado *inductivo* y el *transductivo*. Dado un conjunto de entrenamiento $\{\vec{x}_i, y_i\}_{i=1}^{l}$, $\{\vec{x}_j\}_{j=l+1}^{l+u}$, el aprendizaje semi-supervisado inductivo considera una función $f : X \to Y$ tal que f se espera sea un buen predictor de datos futuros más allá de $\{\vec{x}\}_{j=l+1}^{l+u}$. Por otra parte, el aprendizaje semi-supervisado transductivo entrena una función $f : X^{l+u} \to Y^{l+u}$ tal que f se espera sea un buen predictor de datos no etiquetados $\{\vec{x}_j\}_{j=l+1}^{l+u}$.

II.3. Aprendizaje semi-supervisado

El aprendizaje semi-supervisado es el estudio de cómo los datos etiquetados en conjunto con los no etiquetados pueden cambiar el comportamiento de aprendizaje. En el aprendizaje no supervisado (e.g., agrupación, detección de valores distintos) los datos no están etiquetados, mientras que en el supervisado (e.g., clasificación, regresión) todos los datos tienen etiqueta. El objetivo del aprendizaje semi-supervisado es aprovechar la combinación de estas dos técnicas para cambiar el comportamiento de aprendizaje y

mejorar estos algoritmos, es por ello que es de gran interés en el aprendizaje automático y los datos (Zhu y Goldberg, 2009b).

En esta tesis, se utilizan algoritmos basados en grafos para el aprendizaje semi-supervisado. Este tipo de algoritmos, se basan en construir un grafo cuyos nodos son puntos de datos (etiquetados o no etiquetados) y las aristas representan similitudes entre los puntos. Las etiquetas conocidas se utilizan para propagar información a través del grafo para etiquetar todos los nodos.

Sea $G = (V, E)$ un grafo que representa la "geometría" de los datos, cuyo conjunto de nodos $V = \{1, \ldots, n\}$ representa el conjunto de datos de entrenamiento y el conjunto de aristas E representa las similitudes entre los datos. Estas similitudes se describen a través de una matriz de pesos $\mathbf{W} : \mathbf{W}_{ij}$, la cual es diferente de cero sí y sólo sí x_i y x_j son "vecinos"; i.e., la arista $(i, j) \in E$ ponderada por $\mathbf{W}_{ij}$. La matriz de pesos $\mathbf{W}$ puede ser una matriz de k vecinos más cercanos, tal que $\mathbf{W}_{ij} = 1$ sí y sólo sí x_i se encuentra entre los k vecinos más cercanos de x_j (o viceversa). Generalmente, la matriz de pesos se define por un kernel Gaussiano con un ancho de σ, tal que

$$\mathbf{W}_{ij} = e^{-\frac{||x_i - x_j||^2}{2\sigma^2}}. \tag{1}$$

II.3.1. Algoritmo LabelPropagation

Una idea simple para el aprendizaje semi-supervisado es propagar las etiquetas a través del grafo G. La propagación inicia con los nodos $1, 2, \ldots, l$ con etiquetas conocidas (1 ó -1) y continúa con los nodos l+1, $\ldots, n$. Cada nodo propaga su etiqueta a sus vecinos y este proceso se repite hasta alcanzar a todos los nodos y lograr una *convergencia*. El algoritmo de LabelPropagation propuesto por Zhu y Ghahramani (2002) se basa en

esta idea. En este algoritmo, las etiquetas de los datos se denotan como $\hat{Y} = (\hat{Y}_l, \hat{Y}_u)$, donde $\hat{Y}_l$ puede ser distinto de las etiquetas originales $Y_l = (y_1, \ldots, y_l)$. Aquí, $\hat{Y}_l$ está restringida a ser igual a Y_l. El Pseudo-código 1 muestra una versión del algoritmo de LabelPropagation de acuerdo a la notación propuesta en (Chapelle *et al.*, 2006).

Pseudo-código 1 LabelPropagation

Entrada: El grafo $G(V, E)$ y las etiquetas Y_l
Salida : Las etiquetas $\hat{Y}$

1: Calcula la matriz de pesos $\mathbf{W}$ utilizando la Ecuación 1
2: Genera una matriz diagonal $\mathbf{D}$ donde $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$
3: Inicializa $Y^{(0)} = (y_1, \ldots, y_l, 0, 0, \ldots, 0)$
4: **repeat**
5: $\quad Y^{(t+1)} = \mathbf{D}^{-1}\mathbf{W}\hat{Y}^{(t)}$
6: $\quad Y_l^{(t+1)} = Y_l$
7: **until** alcanzar la convergencia a $\hat{Y}^{(\infty)}$
8: Etiqueta cada punto x_i con el signo de $\hat{y}_i^{(\infty)}$

II.3.2. Algoritmo LabelSpreading

Otro algoritmo de propagación de etiquetas es el LabelSpreading, propuesto por Zhou *et al.* (2004). En cada paso de este algoritmo, un nodo i recibe una contribución de cada uno de sus vecinos j (ponderado por el peso normalizado de la arista (i, j), y una pequeña contribución adicional contemplado en su valor inicial. El Pseudo-código 2 muestra una versión del algoritmo LabelSpreading siguiendo la notación de (Chapelle *et al.*, 2006).

Intuitivamente, el algoritmo LabelSpreading consiste en inferir instancias desconocidas a partir de las instancias conocidas y de esta manera difundir iterativamente las etiquetas por medio de una suma ponderada de sus k vecinos más cercanos.

Pseudo-código 2 LabelSpreading

Entrada: El grafo $G(V, E)$ y las etiquetas Y_l
Salida : Las etiquetas $\hat{Y}$

1: Calcula la matriz de pesos $\mathbf{W}$ utilizando la Ecuación 1 para $i \neq j$ (y $W_{ii} = 0$)
2: Genera una matriz diagonal $\mathbf{D}$ donde $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$
3: Inicializa $Y^{(0)} = (y_1, \ldots, y_l, 0, 0, \ldots, 0)$
4: Elije un parámetro $\alpha \; \epsilon \; [0, 1)$
5: **repeat**
6: $\quad Y^{(t+1)} = \alpha \mathcal{L} Y^{(t)} + (1 - \alpha) Y^{(0)}$
7: **until** alcanzar la convergencia a $\hat{Y}^{(\infty)}$
8: Etiqueta cada punto x_i con el signo de $\hat{y}_i^{(\infty)}$

II.3.3. Algoritmo Self-Training

Adicionalmente a los algoritmos de aprendizaje semi-supervisado basados en grafos (LabelPropagation y LabelSpreading), existen otros enfoques que permiten hacer uso de algoritmos tradicionales de aprendizaje supervisado para completar la información del conjunto de datos. Uno de ellos es el algoritmo Self-Training (o de auto-aprendizaje). Este algoritmo se caracteriza por el hecho de que el proceso de aprendizaje utiliza predicciones para "enseñarse a sí mismo" de forma inductiva o transductiva. La idea principal consiste en generar un modelo utilizando únicamente los registros etiquetados en el conjunto de datos, y después utilizar este modelo para predecir las etiquetas de los registros no etiquetados. El Pseudo-código 3 describe al algoritmo Self-Training.

Pseudo-código 3 Self-Training

Entrada: Registros etiquetados $\{\vec{x}_i, y_i\}_{i=1}^{l}$ y no etiquetados $\{\vec{x}_j\}_{j=l+1}^{l+u}$ dentro del conjunto de datos
Salida : Todo el conjunto de datos etiquetado

1: Sea $L = \{\vec{x}_i, y_i\}_{i=1}^{l}$ y $U = \{\vec{x}_j\}_{j=l+1}^{l+u}$
2: **repeat**
3: $\quad$ Entrena f a partir de L utilizando un algoritmo de aprendizaje supervisado
4: $\quad$ Aplica f al conjunto de registros no etiquetados U
5: $\quad$ Remueve un subconjunto S de U y agrega $\{(\vec{x}, f(\vec{x}))|\vec{x} \in S\}$ a L
6: **until** todos los registros del conjunto de datos estén etiquetados

II.4. Corpus de información paralingüística en niños

Son pocos los trabajos de investigación que tratan el reconocimiento automático de fenómenos paralingüísticos en niños. Shobaki *et al.* (2000) proponen un corpus de datos de discursos de niños, denominado el OGI Kids Speech. Este corpus se compone de audios de discursos preparados y espontáneos de 1100 niños desde preescolar hasta secundaria. Sobre este corpus de datos, se presentan algunos modelos para reconocer nuevas palabras que no se encuentran en el conjunto de entrenamiento. Por ejemplo, Safavi *et al.* (2014) proponen un modelo para la identificación automática de grupos de edades de niños, considerando de 5 a 9, de 9 a 13, y de 13 a 16 años. Otro corpus de datos orientado al reconocimiento del habla para niños es el realizado por Andreas *et al.* (2003). Ellos desarrollan un sistema de reconocimiento para mejorar las habilidades básicas de lectura y comprensión en escuelas públicas. Su corpus consiste de una colección de audios y vídeos de 663 niños desde preescolar hasta primaria. Por otra parte, Cucchiarini y Van hamme (2013) proponen el Jasmin Speech Corpus, un corpus de datos obtenido de personas nativas de Alemania y de Países Bajos, y que está enfocado principalmente para niños y personas de la tercera edad. En este corpus, se consideran diferentes grupos de edad y lenguajes maternos distintos. Sin embargo, todos estos corpus carecen de interacciones genuinas que incluyan anotaciones de fenómenos paralingüísticos (Pérez-Espinosa *et al.*, 2018b).

Capítulo III

Metodología

En este capítulo se propone un método para el reconocimiento de fenómenos paralingüísticos en niños utilizando técnicas de aprendizaje semi-supervisado. Este método se compone principalmente de cuatros fases: el pre-procesamiento del corpus, la extracción de características, el balanceo de la distribución y la construcción de los modelos de clasificación (ver Figura 2). La fase de pre-procesamiento del corpus consiste de la estructuración de los archivos de audio y las etiquetas, calculando el nivel de confiabilidad y eliminando la redundancia. Por otro lado, la fase de extracción de características trata de la obtención de información asociada al procesado digital de las señales de audio y su discretización. La fase de balanceo de la distribución consiste en equilibrar el número de casos para cada clase con el fin de mejorar el proceso de clasificación. Finalmente, la fase de construcción de los modelos de clasificación trata de la aplicación y configuración de los algoritmos de aprendizaje supervisado y semi-supervisado. Con la finalidad de poner en contexto la integración de estas cuatro fases, a continuación se describe el corpus de datos que se utilizó en el presente estudio.

III.1. Corpus de datos

El presente trabajo utiliza el corpus de datos propuesto en (Pérez-Espinosa *et al.*, 2020), el cuál contiene anotaciones de fenómenos paralingüísticos de 174 niños hispanohablantes entre seis y once años de edad. De acuerdo a Pérez-Espinosa *et al.* (2020), la estrategia que se utilizó para recabar los datos fue mediante un juego entre

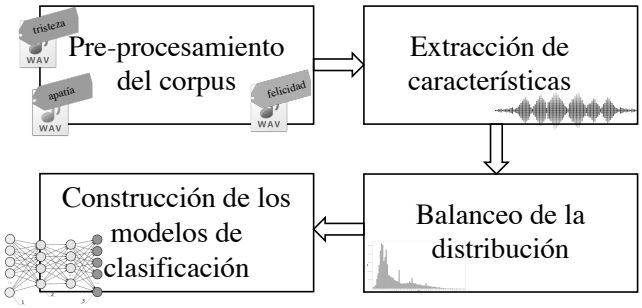


Figura 2. Fases del método propuesto para el reconocimiento de fenómenos paralingüísticos utilizando técnicas de aprendizaje semi-supervisado.

niños y robots, donde cada niño tenía que guiar al robot por medio de la voz para entrar a estaciones y recoger dulces hasta llegar a una meta. Había dos robots, uno bueno y otro malo para seguir indicaciones. El experimento se efectuó en un escenario de Mago de Oz, donde las reacciones afectivas en el habla de los participantes fueron inducidas por medio de robots sociales con el propósito de que éste creará un vínculo social con el niño y facilitara la comodidad e interacción en las actividades diarias. Después de grabar las interacciones, un conjunto de personas (especialistas en lingüística y psicología) etiquetaron cada audio con base en emociones y actitudes de los niños en la interacción con cada robot. Como una primera aplicación, Martínez-Miranda *et al.* (2018) proponen el diseño de robots sociales dirigidos a la población infantil, con el fin de analizar respuestas acústicas afectivas de los niños al interactuar con un robot basado en la comunicación por voz.

El corpus de datos consiste de 30,361 audios de niños, los cuales fueron etiquetados con base en cinco aspectos paralingüísticos: emociones, comportamientos, estado mental, frases y otros fenómenos paralingüísticos en general. Estos aspectos fueron detectados por un equipo de etiquetadores con conocimiento y experiencia en el tema.

A continuación, se describen las fases que integran el método propuesto para el

reconocimiento de fenómenos paralingüísticos.

III.2. Pre-procesamiento del corpus

Esta fase consiste en la generación de tablas estructuradas con los meta-datos de los audios del corpus y sus etiquetas. También aquí se calcula el nivel de confiabilidad de las etiquetas de cada audio o instancia y se discriminan algunos de ellos cuyo segmento fue corrompido o cuya longitud de segmento no es la adecuada. Toda esta información, concentrada originalmente en dos hojas de cálculo, se procesó a través del lenguaje de programación Python con los módulos de pandas y ast.[1] En particular, se construyeron cinco tablas, cada una representando un aspecto paralingüístico de los niños: (1) emociones, (2) comportamientos, (3) estado mental, (4) frases y (5) otros fenómenos paralingüísticos.

La tabla de emociones se compone de ocho clases: desprecio, tristeza, miedo, enojo, sorpresa, felicidad, neutral y ninguno. Las seis primeras, respectivamente, representan las emociones percibidas durante la interacción con los niños. De las otras dos restantes, una de ellas hace referencia a un estado neutral de la emoción, y la otra corresponde al conteo de los audios en los cuales los etiquetadores no detectaron ninguna emoción de las especificadas. La tabla de comportamiento consiste de cuatro clases: entusiasmo, inseguridad, apatía y seguridad, las cuales describen comportamientos contrastantes que presentaron los niños al realizar las diversas actividades. La tabla de estado mental consiste de cuatro clases: confusión, frustración, incertidumbre y ninguno. Similarmente a la tabla de emociones, la clase ninguno denota cuando los etiquetadores no detectaron

[1] La documentación oficial de los módulos de pandas y ast (abstract syntax trees) son: https://pandas.pydata.org y https://docs.python.org/3/library/ast.html.

ningún estado mental de los especificados. La tabla de frases consiste de siete clases. Seis de ellas representan diferentes oraciones que los niños expresaron, y una clase restante en la que los etiquetadores determinaron que el audio no contenía ninguna de las seis oraciones especificadas. Finalmente, la tabla de fenómenos paralingüísticos consiste de siete clases: hyper-articulación, habla a sí mismo, pausa llena, alargamiento de sílaba, grito, reinicio y ninguno. Las seis primeras, respectivamente, representan los diferentes aspectos paralingüísticos articulados en los audios, y la última que denota cuando los etiquetadores no encontraron alguno.

En el presente trabajo, se determinó un *nivel de confiabilidad* a las etiquetas propuestas para cada uno de los audios del corpus. Originalmente, los audios correspondientes a los aspectos paralingüísticos (1), (2) y (3) tienen de una a doce evaluaciones por parte de algunos de los 68 etiquetadores, mientras que los audios de (4) y (5) tienen de cinco a ocho evaluaciones. La etiqueta final para cada audio se asignó con base en la regla de la mayoría, aunque con una confiabilidad asignada. Dicha confiabilidad representa el porcentaje de votos para la etiqueta específica.

Se tiene un total de 30,631 registros originales etiquetados bajo los aspectos (1), (2) y (3). Además, se tiene un total de 10,747 registros etiquetados bajo los los aspectos (4) y (5). Durante el pre-procesamiento, se observó que existían diversos registros duplicados en la hojas de cálculo originales, por lo que se eliminó la redundancia. Algunos de esos registros corresponden a registros de control, los cuales permitían saber si el etiquetador estaba haciendo bien su trabajo, es decir, daba una idea de qué tan bien etiquetados estaban esos registros puesto que era fácil determinar a qué clase pertenecían. Sin embargo, archivos distintos a los de control también estaban duplicados. Los registros de control, tienen la característica de que el nombre del archivo no finaliza con un número. Así que se procedió a eliminar los archivos de control y a seleccionar, de los

archivos duplicados, a aquellos que tuvieran el mayor nivel de consenso.

Por otro lado, de los 30,631 registros, se tiene un total de 29,174 de valores únicos (considerando archivos de datos y de control); mientras que de los 10,747, quedan 10,553. De los aspectos (4) y (5) no existen archivos de control. De los aspectos (1), (2) y (3), se tienen 28,770 registros no duplicados, mientras que de los aspectos (4) y (5) se tienen 10,363. (En este conteo no se incluyen los archivos de datos y de control duplicados.)

Por otra parte, de los registros duplicados, se tienen 1,861 registros para los aspectos (1), (2) y (3). De ese total, se tienen 795 valores únicos. De esos valores únicos, 137 son de control y 658 son registros de datos. De esos registros, se eligieron aquellos que tuvieran el mayor nivel de confiabilidad (que eran 404 registros) y se agregaron a los 28,770 registros no duplicados originalmente. Respecto a los aspectos (4) y (5), todos los registros duplicados son de control. Son 190 registros que, repetidos, son un total de 384 registros. De este total, se eliminaron todos las registros de control y sus duplicados.

Las figuras 3 y 4 muestran una gráfica de distribución de los registros originales después del pre-procesamiento para los aspectos (1), (2) y (3), así como para los aspectos (4) y (5), respectivamente. La Tabla I muestra la distribución final de los audios del corpus considerando las etiquetas de mayor nivel de confiabilidad de los cinco aspectos paralingüísticos después del pre-procesamiento.

III.3. Extracción de características

Se realizó la extracción de características de los audios con ayuda del software openSmile (Eyben *et al.*, 2010). Este fue utilizado para el reconocimiento de emociones por el proyecto openEAR, es por ello que openSmile tiene varios archivos de configuración

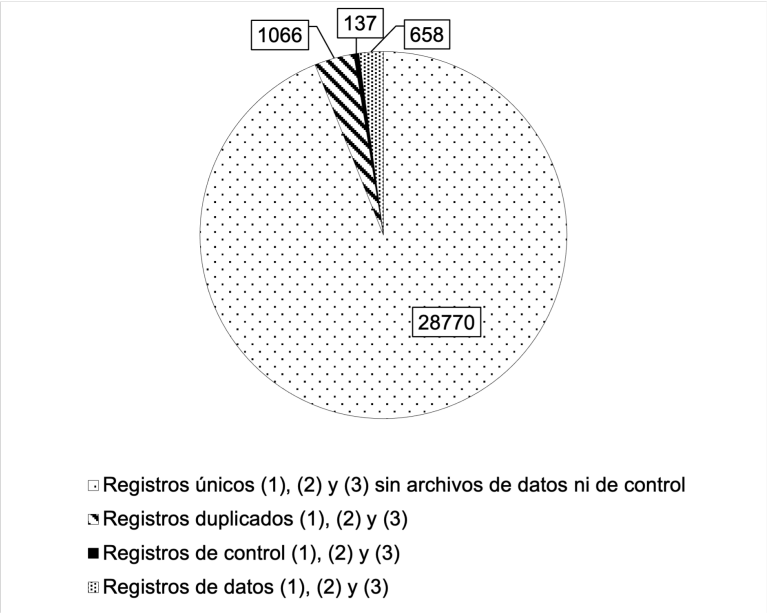


Figura 3. Gráfica de distribución de los 30,631 registros originales etiquetados bajo los aspectos (1), (2) y (3).

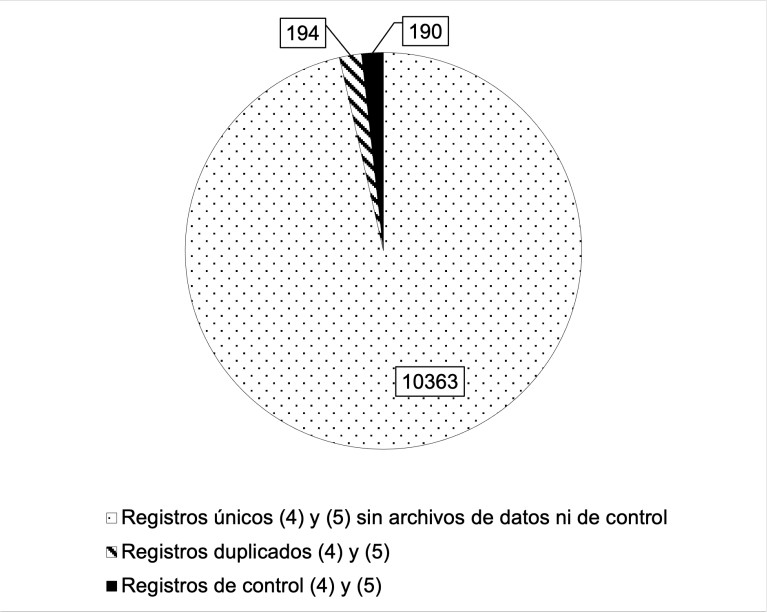


Figura 4. Gráfica de distribución de los 10,747 registros originales etiquetados bajo los aspectos (4) y (5).

Tabla I. Distribución de los audios del corpus con base en las etiquetas de los cinco aspectos paralingüísticos.

Aspectos	Etiquetas	# Audios
Emociones (1)	desprecio	251
	tristeza	725
	miedo	796
	enojo	1017
	sorpresa	2352
	felicidad	6828
	neutral	16140
	ninguno	1065
Comportamiento (2)	entusiasmo	2662
	inseguridad	2807
	apatía	3293
	seguridad	20412
Estado mental (3)	confusión	993
	frustración	1090
	incertidumbre	2414
	ninguno	24677
Frases (4)	yupimi mamá me va a comprar el juguete que yo quiero	102
	¡guacala! ¡hay una cucaracha en mi sopa!	124
	guau!¡llegué a casa y me tenían una sorpresa!	132
	no puede ser! mi hermana me rompió mi juguete favorito	135
	uuuyvi una sombra en mi cuarto	156
	quiero llorara mi perrito lo atropelló un coche	166
	ninguno	9738
Fenómenos paralingüísticos (5)	hyper-articulación	7
	habla a sí mismo	48
	pausa llena	122
	alargamiento de sílaba	268
	grito	349
	reinicio	415
	ninguno	9344

con diferentes características paralingüísticas. Para la extracción de características de los audios se hizo uso de los siguientes archivos de configuración: `IS09_emotion.conf`, `IS10_paraling.conf`, `IS11_speaker_state.conf` y `emo_large.conf`.

El archivo de configuración `IS09_emotion.conf` esta definido por el conjunto de características de INTERSPEECH 2009 Emotion Challenge (Schuller *et al.*, 2009). Este conjunto de características consiste de 16 descriptores de bajos nivel tales como: la tasa de cruces por cero (ZCR, por sus siglas en inglés) de la señal de tiempo, el valor cuadrático medio (RMS, por sus siglas en ingles), la frecuencia de tono normalizada a 500 Hz, la relación armónico-ruido (HNR, por sus siglas en inglés) y los coeficientes cepstrales en las frecuencias de Mel (MFCC, por sus siglas en inglés). Adicionalmente, se incluyen las funcionales de: media, desviación estándar, curtosis, asimetría, el valor

mínimo y máximo, la posición relativa y el rango, así como los coeficientes de la regresión lineal con su error cuadrático medio. En conjunto, el vector de características bajo la configuración de `IS09_emotion.conf` tiene un total de 384 atributos.

Por otra parte, el archivo de configuración `IS10_paraling.conf` se compone del conjunto de características propuestas en el INTERSPEECH 2010 Paralinguistic Challenge. Bajo esta configuración, se obtienen 1,582 características acústicas a través de tres etapas basadas en 'fuerza bruta'. Primero, se consideran 38 descriptores de bajo nivel extraídos de 100 cuadros por segundo, con un tipo de ventana variante en tipo y tamaño y suavizada por un filtro pasa-baja con una longitud de ventana de tres cuadros. Después, se obtienen los coeficientes de regresión de primer orden. Finalmente, se agregan 21 funcionales (media, desviación estándar, curtosis, etc.) por instancia. En (Schuller *et al.*, 2010) se encuentran mayores detalles de estas características.

El archivo de configuración `IS11_speaker_state.conf` consiste del conjunto de características del INTERSPEECH 2011 Speaker State Challenge. Este conjunto consta de un total de 4,468 características, los cuales incluyen información conocida como relevante para las tareas de reconocimiento de emociones (Hollien *et al.*, 2001; Dhupati *et al.*, 2010). Estas características se conforman de tres conjuntos de descriptores de bajo nivel. A diferencia de las configuraciones `IS09_emotion.conf` y `IS10_paraling.conf`, en este conjunto de características se incluye una medida de sonoridad derivada del espectro auditivo y el uso de espectros auditivos filtrados RASTA en lugar de los MFCC. Adicionalmente, aquí se considera un conjunto extendido de descriptores espectrales estadísticos tales como entropia, varianza, etc. En (Schuller *et al.*, 2011) se encuentran mayores detalles de estas características.

Finalmente, el archivo de configuración `emo_large.conf` genera el conjunto de características más grande del openSmile. Esta configuración considera un total de

6,552 características numéricas que reflejan datos de tono, varianza, etc. Todas estas características no están documentadas en la herramienta, pero incluye un gran número de características derivadas tales como media, rango, desviación estándar, cuartiles, rango inter-cuartiles, descriptores y sus coeficientes de regresión delta.

Todos estos archivos de configuración se aplicaron al corpus de audios de Pérez-Espinosa *et al.* (2018a). El total de registros después de la etapa de pre-procesamiento y la extracción de características quedan de la siguiente manera. Para los aspectos (1), (2) y (3) se tiene un total de 29,174 registros y para el (4) y (5) 10,553 registros. Observe que existe una diferencia con respecto al número de registros obtenidos después de pre-procesamiento. Esto se debe a que algunos archivos de audio fueron discriminados de acuerdo a la duración del segmento. En este trabajo, se consideran segmentos de audio de aproximadamente 2.4 segundos como mínimo, esto debido a que para cada uno de los aspectos paralingüísticos se estableció un umbral diferente, con el fin de eliminar aquellos audios donde la duración era muy pequeña y además coadyuvara al balanceo de las clases, es por ello que la duración mínima que se discrimino en cada aspecto paralingüístico es diferente.

La Tabla II muestra la distribución de los registros del corpus con base en las etiquetas de los cinco aspectos paralingüísticos posterior a la extracción de características.

Tabla II. Distribución de los audios del corpus con base en las etiquetas de los cinco aspectos paralingüísticos después de la extracción de características.

Aspectos	**Etiquetas**	**# Audios**
Emociones (1)	desprecio	251
	tristeza	725
	miedo	796
	enojo	1017
	ninguno	1065
	sorpresa	2352
	felicidad	6828
	neutral	16140
Comportamiento (2)	entusiasmo	2662
	inseguridad	2807
	apatía	3293
	seguridad	20412
Estado mental (3)	confusión	993
	frustración	1090
	incertidumbre	2414
	ninguno	24677
Frases (4)	yupimi mamá me va a comprar el juguete que yo quiero	102
	¡guacala! ¡hay una cucaracha en mi sopa!	124
	guau!¡llegué a casa y me tenían una sorpresa!	132
	no puede ser! mi hermana me rompió mi juguete favorito	135
	uuuyvi una sombra en mi cuarto	156
	quiero llorara mi perrito lo atropelló un coche	166
	ninguno	9738
Fenómenos paralingüísticos (5)	hyper-articulación	7
	habla a sí mismo	48
	pausa llena	122
	alargamiento de sílaba	268
	grito	349
	reinicio	415
	ninguno	9344

III.4. Balanceo de la distribución de clases

Se observa en la Tabla II que existe un gran desbalanceo entre la cantidad de registros para cada clase, lo que puede impactar negativamente al desempeño de los modelos de clasificación. Un primer enfoque para solucionar esto, fue el utilizar la combinación de diversas técnicas de balanceo tanto de sobre-muestreo (over-sampling) como de sub-muestreo (under-sampling). Las técnicas de sobre-muestro son aquellas que ayudan a muestrear un conjunto de datos originales sobremuestreando la clase minoritaria esto puede llevar a remuestrar muestras no tan valiosas, es decir, aquellas que no alcanzaron un buen nivel de consenso; mientras que por otra parte, las técnicas de sub-muestreo

elimina muestras de la clase mayoritaria, cabe mencionar que al eliminar de manera aleatoria puede suceder que se pierdan muestran con información importante.

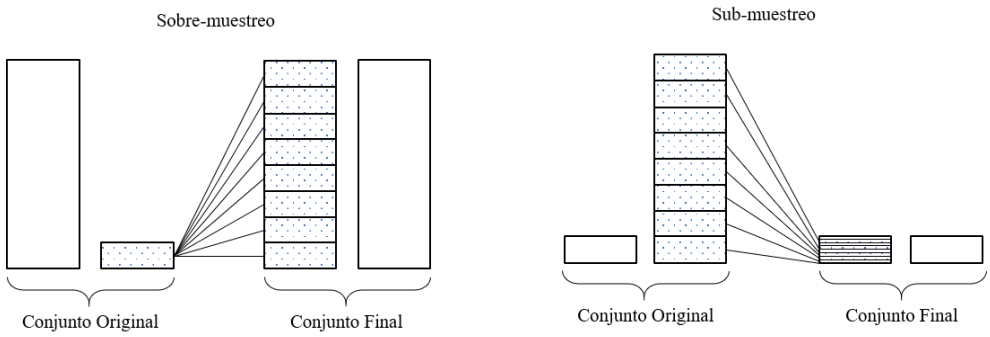


Figura 5. Técnicas de Balanceo

Las técnicas de sobre-muestro que se utilizaron son las siguientes:

O1) SMOTE (synthetic minority over-sampling technique). Esta técnica consiste en agregar datos "sintéticos" a las clases minoritarias a partir de un conjunto de vectores de características elegidos aleatoriamente y considerando sus k vecinos más cercanos. Cada vector característica "sintético", considera la diferencia entre el vector característica elegido y su vecino más cercano; se multiplica esa diferencia por un valor aleatorio entre 0 y 1, y se le suma al vector característica en consideración (Chawla *et al.*, 2002).

O2) Borderline-SMOTE. Esta variante de la implementación de SMOTE soporta el sobre-muestreo multiclase siguiendo un esquema de uno versus el resto. El término "límite" (borderline) hace referencia a que solamente se consideran conjuntos de características cercanos a la frontera de decisión para generar nuevos datos (Han *et al.*, 2005).

Las técnicas de sub-muestro que se utilizaron son las siguientes

U1) All KNN. Esta técnica incluye el uso del algoritmo de aprendizaje de los K vecinos más cercanos (o KNN, por sus siglas en inglés). Para cada vector de características de la clase mayoritaria, se aplica el algoritmo KNN para etiquetar los otros vectores características del conjunto. Si la mayoría de sus K vecinos más cercanos se asignan a las clases minoritarias, entonces el vector característica en cuestión se elimina (Tomek *et al.*, 1976).

U2) Cluster-Centroids. Esta técnica se basa en calcular centroides (o centroids) en agrupamientos (o clusters). Para el cálculo de estos centroides se utiliza el algoritmo de aprendizaje no supervisado K-means, donde K corresponde al número de muestras minoritarias y converge cuando las observaciones se reasignan (Douzas *et al.*, 2018).

U3) NearMiss. Si los vectores de características de dos casos pertenecientes a clases distintas están cerca el uno del otro, esta técnica descarta estos vectores de la clase mayoritaria con la finalidad de aumentar los espacios entre ambas clases y mejorar la clasificación (Mani y Zhang, 2003).

Todas estas técnicas se encuentran implementadas en el módulo imbalanced-learn del scikit-learn de Python.

Las técnicas de sobre-muestreo y sub-muestreo se combinaron para reducir las clases mayoritarias y, al mismo tiempo, re-muestrear las clases minoritarias. Las combinaciones fueron las siguientes: All KNN con SMOTE, AllKNN con Borderline-SMOTE, Cluster-Centroids con Borderline-SMOTE, NearMiss con Borderline-SMOTE y NearMiss con SMOTE. Note que no se realizaron todas las combinaciones, como por ejemplo Cluster-Centroids con SMOTE, debido a que se conoce de antemano que Borderline-SMOTE es una mejor versión SMOTE. Las combinaciones de las técnicas de sobre-muestreo y

sub-muestreo se aplicaron a la distribución de los audios de la Tabla II para cada configuración de extracción de características. Esto debido a que los atributos o características de los audios obtenidas por cada archivo de configuración, proveen resultados distintos para cada combinación de las técnicas de balanceo. La distribución de los audios después del proceso de balanceo se muestra en las tablas III y IV para la configuración `IS09_emotion.conf`, en las tablas V y VI para la configuración `IS10_paraling.conf`, en las tablas VII y VIII para la configuración `IS11_speaker_state.conf`, y en las tablas IX y X para la configuración `emo_large.conf`. Cabe destacar que no todas las combinaciones propuestas funcionaron, principalmente en los conjuntos cuyo número de atributos eran grandes como las obtenidas por los archivos de configuración `IS11_speaker_state.conf` y `emo_large.conf`. Para estos últimos casos, adicionalmente se utilizó una técnica de selección de características para la reducción de dimensionalidad. En este caso se utilizó el algoritmo de umbrales de varianza (o *variance threshold*) para remover todas las características de varianza cero; i.e., se eliminan aquellas que tiene el mismo valor en todas las muestras. Las tablas XI y XII muestran la distribución de los audios aplicando el algoritmo de umbrales de varianza (denotado por la letra V) para la reducción de la dimensionalidad y previo al balanceo de los datos.

Las combinaciones de las técnicas de balanceo se aplican bajo un enfoque de uniformidad de distribución de los niveles de confiabilidad de las etiquetas de los audios. Para ello, se consideran cuatro intervalos de los niveles de confiabilidad: $[0, 0.2)$, $[0.2, 0.4)$, $[0.4, 0.8)$ y $[0.8, 1]$. Los resultados para esta primera versión del trabajo se presentan en la Sección IV.1.

De manera alterna a las combinaciones de las técnicas de muestreo, y como un segundo enfoque, se aplicaron umbrales de discriminación en la longitud de los segmentos de audios, con el fin de lograr un equilibrio en la distribución de las

Tabla III. Distribución de los audios de la configuración `IS09_emotion.conf` con base a las técnicas de balanceo considerando los aspectos de emociones, estado mental y paralingüística.

Emociones							
Etiquetas	# Audios	U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
neutral	16140	6610	6610	6844	6844	6900	6900
felicidad	6828	6836	6836	6836	6836	6836	6836
sorpresa	2352	3057	3057	3057	3057	3057	3057
ninguno	1065	3071	3071	3071	3071	3071	3071
enojo	1017	3055	3055	3055	3055	3055	3055
miedo	796	3055	3055	3055	3055	3055	3055
tristeza	725	3055	3055	3055	3055	3055	3055
desprecio	251	3055	3055	3055	3055	3055	3055
Estado Mental							
Etiquetas	# Audios	U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
ninguno	24677	-	-	1461	1461	1463	1463
incertidumbre	2414	-	-	1461	1461	1461	1461
frustración	1090	-	-	1461	1461	1461	1461
confusión	993	-	-	1465	1465	1465	1465
Paralingüística							
Etiquetas	# Audios	U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
ninguno	9344	-	-	349	349	349	349
reinicio	415	-	-	349	349	349	349
grito	349	-	-	349	349	349	349
alargamiento de sílaba	268	-	-	268	268	268	268
pausa llena	122	-	-	122	122	122	122
habla a sí mismo	48	-	-	122	122	122	122
hyper-articulación	7	-	-	122	122	122	122

Tabla IV. Distribución de los audios de la configuración `IS09_emotion.conf` con base a las técnicas de balanceo considerando los aspectos de frases y comportamiento.

Frases					
Etiquetas	# Audios	U2&O2	U2&O1	U3&O2	U3&O1
ninguno	9738	-	166	-	166
quiero llorar a mi perrito lo atropelló un coche	166	-	166	-	166
¡uuuuy! vi una sombra en mi cuarto	156	-	156	-	156
¡no puede ser! mi hermana me rompió mi juguete favorito	135	-	135	-	135
¡guau!¡llegué a casa y me tenían una sorpresa!	132	-	132	-	132
¡guacala! ¡hay una cucaracha en mi sopa!	124	-	124	-	124
¡yupi! mi mamá me va a comprar el juguete que yo quiero	102	-	102	-	102
Comportamiento					
Etiquetas	# Audios	U2&O2	U2&O1	U3&O2	U3&O1
seguridad	20412	-	2438	-	2438
apatía	3293	-	2337	-	2337
inseguridad	2807	-	2555	-	2555
entusiasmo	2662	-	2287	-	2287

Tabla V. Distribución de los audios de la configuración `IS10_paraling.conf` con base a las técnicas de balanceo considerando los aspectos de emociones, estado mental y paralingüística.

Emociones							
Etiquetas	# Audios	U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
neutral	16140	7231	7231	7032	7032	7098	7098
felicidad	6828	7024	7024	7024	7024	7024	7024
sorpresa	2352	2978	2978	2978	2978	2978	2978
ninguno	1065	2992	2992	2992	2992	2992	2992
enojo	1017	2976	2976	2976	2976	2976	2976
miedo	796	2976	2976	2976	2976	2976	2976
tristeza	725	2976	2976	2976	2976	2976	2976
desprecio	251	2976	2976	2976	2976	2976	2976
Estado Mental							
Etiquetas	# Audios	U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
ninguno	24677	-	-	1461	1461	1467	1467
incertidumbre	2414	-	-	1461	1461	1461	1461
frustración	1090	-	-	1461	1461	1461	1461
confusión	993	-	-	1465	1465	1465	1465
Paralingüística							
Etiquetas	# Audios	U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
ninguno	9344	-	-	349	349	349	349
reinicio	415	-	-	349	349	349	349
grito	349	-	-	349	349	349	349
alargamiento de sílaba	268	-	-	268	268	268	268
pausa llena	122	-	-	122	122	122	122
habla a sí mismo	48	-	-	122	122	122	122
hyper-articulación	7	-	-	122	122	122	122

Tabla VI. Distribución de los audios de la configuración `IS10_paraling.conf` con base a las técnicas de balanceo considerando los aspectos de frases y comportamiento.

Frases					
Etiquetas	# Audios	U2&O2	U2&O1	U3&O2	U3&O1
ninguno	9738	-	166	-	166
quiero llorar a mi perrito lo atropelló un coche	166	-	166	-	166
¡uuuuy! vi una sombra en mi cuarto	156	-	156	-	156
¡no puede ser! mi hermana me rompió mi juguete favorito	135	-	135	-	135
¡guau!¡llegué a casa y me tenían una sorpresa!	132	-	132	-	132
¡guacala! ¡hay una cucaracha en mi sopa!	124	-	124	-	124
¡yupi! mi mamá me va a comprar el juguete que yo quiero	102	-	102	-	102
Comportamiento					
Etiquetas	# Audios	U2&O2	U2&O1	U3&O2	U3&O1
seguridad	20412	-	2438	-	2438
apatía	3293	-	2337	-	2337
inseguridad	2807	-	2555	-	2555
entusiasmo	2662	-	2287	-	2287

Tabla VII. Distribución de los audios de la configuración `IS11_speaker_state.conf` con base a las técnicas de balanceo considerando los aspectos de emociones, estado mental y paralingüística.

Emociones							
Etiquetas	# Audios	U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
neutral	16140	6567	6567	7032	7032	7032	7032
felicidad	6828	7016	7016	7016	7016	7016	7016
sorpresa	2352	2976	2976	2976	2976	2976	2976
ninguno	1065	2992	2992	2992	2992	2992	2992
enojo	1017	2976	2976	2976	2976	2976	2976
miedo	796	2976	2976	2976	2976	2976	2976
tristeza	725	2976	2976	2976	2976	2976	2976
desprecio	251	2976	2976	2976	2976	2976	2976
Estado Mental							
Etiquetas	# Audios	U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
ninguno	24677	-	-	1461	1461	1461	1461
incertidumbre	2414	-	-	1461	1461	1461	1461
frustración	1090	-	-	1461	1461	1461	1461
confusión	993	-	-	1461	1461	1461	1461
Paralingüística							
Etiquetas	# Audios	U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
ninguno	9344	-	-	349	349	349	349
reinicio	415	-	-	349	349	349	349
grito	349	-	-	349	349	349	349
alargamiento de sílaba	268	-	-	268	268	268	268
pausa llena	122	-	-	122	122	122	122
habla a sí mismo	48	-	-	122	122	122	122
hyper-articulación	7	-	-	122	122	122	122

Tabla VIII. Distribución de los audios de la configuración `IS11_speaker_state.conf` con base a las técnicas de balanceo considerando los aspectos de frases y comportamiento.

Frases			
Etiquetas	# Audios	U2&O1	U3&O1
ninguno	9738	166	166
quiero llorara mi perrito lo atropelló un coche	166	166	166
uuuyvi una sombra en mi cuarto	156	156	156
no puede ser! mi hermana me rompió mi juguete favorito	135	135	135
guau!¡llegué a casa y me tenían una sorpresa!	132	132	132
¡guacala! ¡hay una cucaracha en mi sopa!	124	124	124
yupimi mamá me va a comprar el juguete que yo quiero	102	102	102
Comportamiento			
Etiquetas	# Audios	U2&O1	U3&O1
seguridad	20412	2586	2586
apatía	3293	2484	2484
inseguridad	2807	2294	2294
entusiasmo	2662	2318	2318

Tabla IX. Distribución de los audios de la configuración `emo_large.conf` con base a las técnicas de balanceo considerando los aspectos de emociones, estado mental y paralingüística.

Emociones							
Etiquetas	# Audios	U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
neutral	16140	6753	6753	7032	7032	7032	-
felicidad	6828	7016	7016	7016	7016	7016	-
sorpresa	2352	2976	2976	2976	2976	2976	-
ninguno	1065	2992	2992	2992	2992	2992	-
enojo	1017	2976	2976	2976	2976	2976	-
miedo	796	2976	2976	2976	2976	2976	-
tristeza	725	2976	2976	2976	2976	2976	-
desprecio	251	2976	2976	2976	2976	2976	-
Estado Mental							
Etiquetas	# Audios	U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
ninguno	24677	-	-	-	-	1461	-
incertidumbre	2414	-	-	-	-	1461	-
frustración	1090	-	-	-	-	1461	-
confusión	993	-	-	-	-	1461	-
Paralingüística							
Etiquetas	# Audios	U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
ninguno	9344	-	-	349	349	349	349
reinicio	415	-	-	349	349	349	349
grito	349	-	-	349	349	349	349
alargamiento de sílaba	268	-	-	268	268	268	268
pausa llena	122	-	-	122	122	122	122
habla a sí mismo	48	-	-	122	122	122	122
hyper-articulación	7	-	-	122	122	122	122

Tabla X. Distribución de los audios de la configuración `emo_large.conf` con base a las técnicas de balanceo considerando los aspectos de frases y comportamiento.

Frases			
Etiquetas	# Audios	U2&O1	U3&O1
ninguno	9738	166	166
quiero llorara mi perrito lo atropelló un coche	166	166	166
uuuyvi una sombra en mi cuarto	156	156	156
no puede ser! mi hermana me rompió mi juguete favorito	135	135	135
guau!¡llegué a casa y me tenían una sorpresa!	132	132	132
¡guacala! ¡hay una cucaracha en mi sopa!	124	124	124
yupimi mamá me va a comprar el juguete que yo quiero	102	102	102
Comportamiento			
Etiquetas	# Audios	U2&O1	U3&O1
seguridad	20412	2438	2586
apatía	3293	2885	2484
inseguridad	2807	4380	2294
entusiasmo	2662	2287	2318

Tabla XI. Distribución de los audios de la configuración `IS11_speaker_state.conf` después de la reducción de dimensionalidad, considerando los aspectos de emociones, estado mental y comportamiento.

Emociones						
Etiquetas	# Audios	U1&O1&V	U2&O2&V	U2&O1&V	U3&O2&V	U3&O1&V
neutral	16140	6567	7032	7032	7032	7032
felicidad	6828	7016	7016	7016	7016	7016
sorpresa	2352	2976	2976	2976	2976	2976
ninguno	1065	2992	2992	2992	2992	2992
enojo	1017	2976	2976	2976	2976	2976
miedo	796	2976	2976	2976	2976	2976
tristeza	725	2976	2976	2976	2976	2976
desprecio	251	2976	2976	2976	2976	2976
Estado Mental						
Etiquetas	# Audios	U1&O1&V	U2&O2&V	U2&O1&V	U3&O2&V	U3&O1&V
ninguno	24677	-	1461	1461	1461	1461
incertidumbre	2414	-	1461	1461	1461	1461
frustración	1090	-	1461	1461	1461	1461
confusión	993	-	1461	1461	1461	1461
Comportamiento						
Etiquetas	# Audios	U1&O1&V	U2&O2&V	U2&O1&V	U3&O2&V	U3&O1&V
seguridad	20412	-	-	2586	-	2586
apatia	3293	-	-	2484	-	2484
inseguridad	2807	-	-	2294	-	2294
entusiasmo	2662	-	-	2318	-	2318

Tabla XII. Distribución de los audios de la configuración `emo_large.conf` después de la reducción de dimensionalidad, considerando los aspectos de emociones y comportamiento.

Emociones				
Etiquetas	# Audios	U2&O1&V	U3&O2&V	U3&O1&V
neutral	16140	-	7032	7032
felicidad	6828	-	7016	7016
sorpresa	2352	-	2976	2976
ninguno	1065	-	2992	2992
enojo	1017	-	2976	2976
miedo	796	-	2976	2976
tristeza	725	-	2976	2976
desprecio	251	-	2976	2976
Comportamiento				
Etiquetas	# Audios	U2&O1&V	U3&O2&V	U3&O1&V
seguridad	20412	2438	-	2438
apatia	3293	2885	-	2885
inseguridad	2807	4380	-	4380
entusiasmo	2662	2287	-	2287

clases. Cabe resaltar que la mayoría de los audios tiene una duración promedio de 1.9 segundos, pero pueden ser muy variados dependiendo del aspecto paralingüístico a tratar. Como regla de un límite superior, se discriminaron todos los audios cuya duración fuera mayor a 30 segundos; mientras que como límite inferior, se eliminaron los audios cuya duración era menor a 2.4 segundos en promedio. El limite inferior estuvo sujeto a la cantidad de registros y se aplicó únicamente a las clases mayoritarias. Las Tablas XIII y XIV muestran la distribución de los registros después de aplicar los umbrales de discriminación con base en la longitud de los segmentos de audio. Estos resultados fueron los mismos para las tres configuraciones `IS09_emotion.conf`, `IS10_speaker_state.conf`, `IS11_paraling.conf` y `emo_large.conf`.

Tabla XIII. Distribución de los audios después de eliminar los segmentos $\sim<= 2.4s$ y $> 30s$, considerando los aspectos de emociones, comportamiento y estado mental.

Emociones	
Etiquetas	# Audios
neutral	1264
sorpresa	1180
felicidad	1167
ninguno	1064
enojo	1017
miedo	796
tristeza	725
desprecio	251
Comportamiento	
Etiquetas	# Audios
seguridad	2813
inseguridad	2807
entusiasmo	2662
apatía	2590
Estado Mental	
Etiquetas	# Audios
ninguno	1578
incertidumbre	1563
frustración	1090
confusión	993

Tabla XIV. Distribución de los audios después de eliminar los segmentos $\leq 2.4s$ y $> 30s$, considerando los aspectos de frases y paralingüística.

Frases	
Etiquetas	# Audios
ninguno	199
quiero llorara mi perrito lo atropelló un coche	166
uuuyvi una sombra en mi cuarto	156
no puede ser! mi hermana me rompió mi juguete favorito	135
guau!¡llegué a casa y me tenían una sorpresa!	132
¡guacala! ¡hay una cucaracha en mi sopa!	124
yupimi mamá me va a comprar el juguete que yo quiero	102
Paralingüística	
Etiquetas	# Audios
grito	286
alargamiento de sílaba	268
reinicio	256
ninguno	254
pausa llena	122
habla a sí mismo	48
hyper-articulación	7

III.5. Modelos de clasificación

Esta sección describe la generación de los modelos de clasificación, los cuales se entrenaron utilizando registros etiquetados bajo diferentes niveles de confiabilidad. Para ello, fue necesario "validar" los registros con etiquetas menos confiables a partir de modelos de predicción obtenidos considerando únicamente registros con etiquetas confiables. Este proceso de validación consiste en utilizar técnicas de aprendizaje semi-supervisado, para conocer cuáles de los registros que utilizan etiquetas menos confiables coinciden con las etiquetas predichas por los modelos.

Durante el desarrollo del presente trabajo, se consideraron diversos umbrales para decidir cuáles registros inicialmente eran confiables o no. Sin embargo, algunos resultados experimentales preliminares de los modelos mostraron ser mejores cuando se consideraban registros en el que se alcanzó consenso; i.e., aquellos cuyo nivel de confiabilidad es exactamente uno. De estos registros con etiquetas confiables, se definieron dos nuevos conjuntos, el conjunto entrenamiento y el de validación. Es

precisamente el conjunto entrenamiento, el que se utiliza para generar los modelos de predicción que permiten validar los registros con etiquetas menos confiables. Finalmente, se considera un modelo de clasificación que considera los registros del conjunto entrenamiento junto con los registros con las etiquetas validadas, y se prueba su desempeño utilizando el conjunto validación. En la Figura 6 se muestra un esquema general de este proceso.

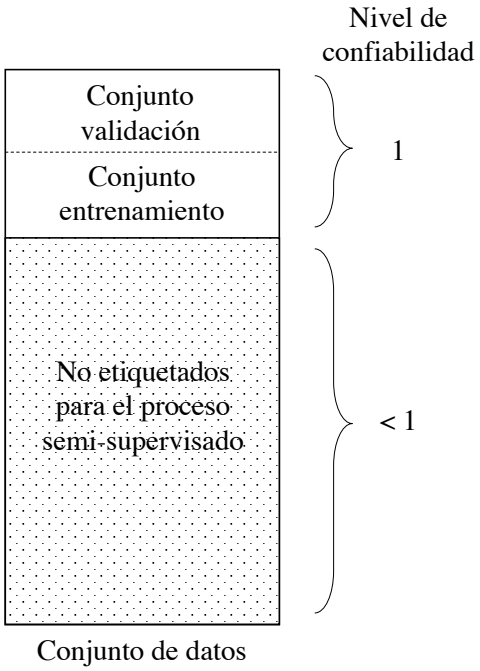


Figura 6. Partición de datos para el entrenamiento semi-supervisado.

III.5.1. Balanceo de los registros con etiquetas asignadas por consenso

Una vez que se discriminaron los registros en los que se alcanzó consenso, fue necesaria una segunda etapa de balanceo en la distribución de las clases. Esto debido a que el número de registros varía acorde a su nivel de confiabilidad. Un ejemplo de ello

se muestra en la Tabla XV. La columna A de la tabla muestra la distribución de los registros donde se alcanzó consenso para el aspecto paralingüístico de emociones. Allí se observa que dicha distribución no está balanceada. Para resolver esto, se consideró únicamente un porcentaje de la clase mayoritaria (alrededor de un 30 %), que en este ejemplo fue la clase con etiqueta "ninguno". El resultado de este muestreo de la clase mayoritaria se presenta en la columna B de la Tabla XV. Posteriormente, se utilizó la técnica de sobre-muestreo SMOTE para generar nuevos registros en el resto de las clases. El total de nuevos registros por clase es igual al porcentaje de la clase mayoritaria (ver la columna C de la Tabla XV). Finalmente, se consideran los nuevos registros junto con los registros originales y se obtiene una distribución de clases. La columna D de la Tabla XV muestra la distribución final de registros por clase.

Tabla XV. Segunda etapa de balanceo de la distribución de los audios de emociones cuyas etiquetas fueron asignadas mediante consenso.

Etiquetas	A	B	C	D
felicidad	74	74	200	274
ninguno	627	200	200	627
sorpresa	16	16	200	216
neutral	57	57	200	257
enojo	19	19	200	219
miedo	14	14	200	214
tristeza	42	42	200	242
desprecio	18	18	200	218

Los conjuntos de entrenamiento y de validación se conforman de los registros resultantes de la segunda etapa de balanceo. Inicialmente, se considera un total de 150 registros por clase para el conjunto entrenamiento y el resto para el conjunto validación. Este criterio se aplica para todos los modelos de los aspectos paralingüísticos. El resto de los registros, cuyas etiquetas se asignaron con un nivel de confiabilidad menor a uno, se validan en sus etiquetas utilizando dicho conjunto de entrenamiento inicial.

Siguiendo el ejemplo del aspecto paralingüístico de emociones, la Tabla XVI

muestra la división del conjunto entrenamiento y el conjunto validación de los registros etiquetados por consenso. La Tabla XVII muestra la distribución por clases del conjunto de audios cuyo nivel de confiabilidad es estrictamente menor que uno, para el mismo ejemplo del aspecto paralingüístico de emociones.

Tabla XVI. Distribución de audios de emociones en los conjuntos de entrenamiento y de validación.

Etiquetas	Conjunto entrenamiento	Conjunto validación
felicidad	150	124
ninguno	150	477
sorpresa	150	66
neutral	150	107
enojo	150	69
miedo	150	64
tristeza	150	92
desprecio	150	68

Tabla XVII. Distribución de audios cuyas etiquetas tiene un nivel de confiabilidad < 1.

Etiquetas	Registros cuyas etiquetas se asignaron sin consenso
felicidad	1093
ninguno	437
sorpresa	1164
neutral	1207
enojo	998
miedo	782
tristeza	683
desprecio	233

III.5.2. Validación de etiquetas a través de técnicas semi-supervisadas

Una vez que se define el conjunto de entrenamiento inicial, se utilizan algoritmos de aprendizaje semi-supervisado para validar las etiquetas de los registros cuyo nivel de confiabilidad es menor que uno. Los algoritmos que se utilizaron en el presente trabajo fueron LabelPropagation, LabelSpreading y Self-Training (ver Sección II.3).

Los primeros dos algoritmos, respectivamente, crean un grafo de similitud de los datos de entrada para propagar las etiquetas de los audios etiquetados con un mayor nivel de confiabilidad y mejorar las etiquetas de los audios con un menor nivel (ver Figura 7). El algoritmo Self-Training utiliza modelos de aprendizaje supervisado para estimar la etiqueta de los registros no etiquetados. Se utilizó una versión de la implementación de estas técnicas disponible en la librería scikit-learn (Pedregosa *et al.*, 2011) de Python, dentro del módulo sklearn.semi_supervised.[2]

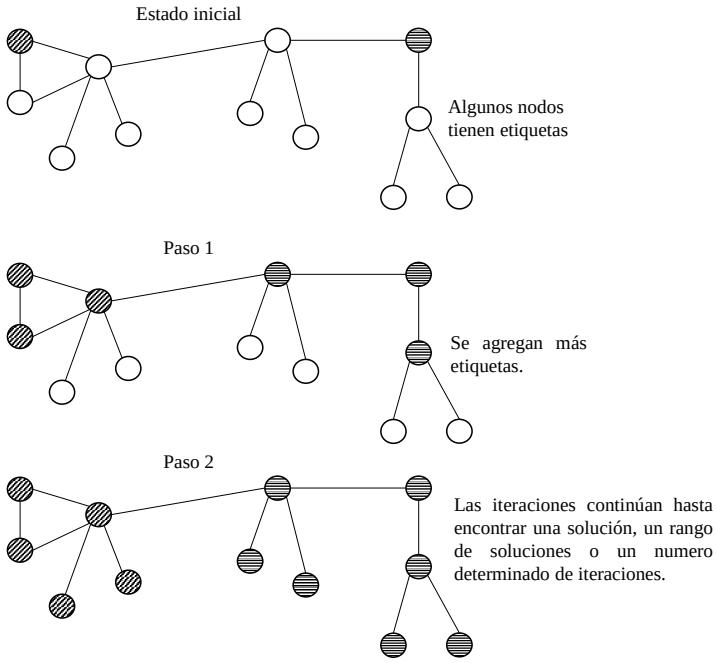


Figura 7. Grafo de similitud de las técnicas de aprendizaje semi-supervisado LabelPropagation y LabelSpreading

[2]La documentación oficial se encuentra en: `https://scikit-learn.org/stable/modules/semi_supervised.html`.

Los algoritmos LabelPropagation y LabelSpreading difieren en la conformación de la matriz de adyacencias que define el grafo de similitud. Esta matriz bidimensional contiene las medidas de similitud entre pares de elementos de la secuencia (Rafii y Pardo, 2012). LabelPropagation construye la matriz de similitud a partir de los vectores características sin ningún tratamiento adicional, mientras que LabelSpreading lo hace minimizando una función de perdida permitiendo una mayor resistencia al ruido. Ambos algoritmos manejan dos tipos de núcleos o *kernels* que impactan en la escalabilidad y eficiencia de sus resultados de re-etiquetado: el *kernel* rbf (*radial-basis function*) que se basa en funciones radiales y el *kernel* KNN (K-*Nearest Neighbors*) que se basa en la técnica de K vecinos más cercanos. Las pruebas experimentales preliminares mostraron un desempeño pobre y lento utilizando el *kernel* rbf, e incluso en algunos casos, no se lograba la convergencia. Por lo tanto, en este trabajo se optó por utilizar el *kernel* KNN para estos algoritmos.

III.5.3. Modelos de clasificación de aprendizaje supervisado

Una vez que se validaron los audios a través del proceso de re-etiquetado, se generaron los modelos de clasificación utilizando los siguientes algoritmos de aprendizaje supervisado: la clasificación de vector de soporte lineal (LSVC), el algoritmo de K vecinos más cercanos (KNN), bosques aleatorios (RF), árboles de decisión (DT), redes neuronales artificiales (NN), naive Bayes (NB), el algoritmo de impulso adaptativo (AB) y el análisis discriminativos lineal (LDA) y el cuadrático (QDA). Se utilizó la implementación de todos estos algoritmos incluida en la librería del scikit-learn de Python[3].

[3]La documentación oficial se encuentra en: `https://scikit-learn.org/stable/supervised_learning.html`

III.6. Modelos dependientes / independientes del hablante

Los modelos de clasificación de la Sección III.5 podrían tener cierta dependencia del hablante, ya que muestras de la voz del mismo hablante se encuentran tanto en el conjunto entrenamiento como en el conjunto validación. Esto puede dar lugar a un sesgo en los resultados de la eficiencia de los modelos de clasificación, ya que puede darse el caso que tales modelos consideren, además de los aspectos paralingüísticos, información relacionada a las particularidades acústicas de la voz del hablante.

Para evitar sesgos en los resultados, se propone la generación de otros modelos de clasificación que sean totalmente independientes del hablante. La finalidad de los modelos independientes del hablante, es almacenar patrones generales de un grupo de usuarios más que de uno en particular. En estos nuevos modelos, ningún hablante que aparece en el conjunto entrenamiento estará en el conjunto validación. Con este diseño se determina si el modelo debe entrenarse para cada usuario, o si es posible hacer la independencia del hablante.

De forma similar a los modelos dependientes del hablante, los modelos independientes consideran los registros con mejor nivel de consenso para dividir los registros cuyas etiquetas alcanzaron consenso y aquellos cuyo nivel de confiabilidad es menor que uno (ver Figura 8). Los pasos de tratamiento son los mismos: se lleva a cabo el balanceo de los registros con etiquetas asignadas por consenso, se validan las etiquetas a través de técnicas semi-supervisadas y se generan los modelos de clasificación a través del aprendizaje supervisado.

Finalmente, los modelos de clasificación tanto dependientes como independientes del hablante, se implementaron utilizando los conjuntos de datos balanceados a través

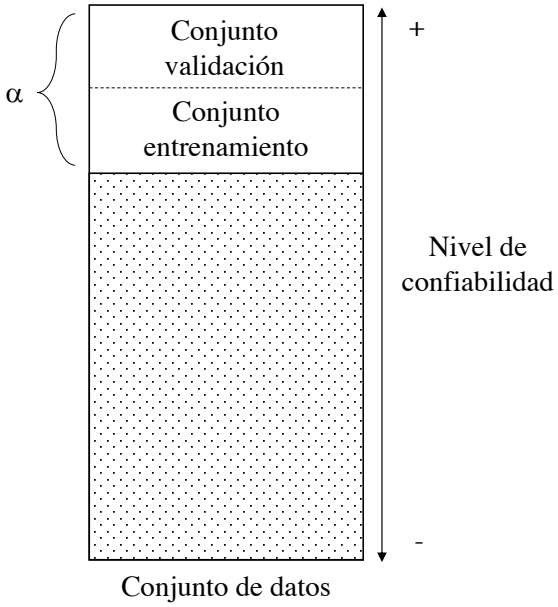


Figura 8. Diagrama del modelo independiente de datos.

del enfoque de los umbrales de discriminación en la longitud de los segmentos de audio (de las Tablas XIII y XIV). Los resultados de estos modelos se presentan en las secciones IV.2 y IV.3 del siguiente capítulo.

Capítulo IV

Resultados experimentales

En este capítulo se presentan los resultados de un estudio experimental comparativo acerca de la eficiencia de los modelos de clasificación propuestos, tanto dependientes como independientes del hablante. Para comparar el desempeño de estos modelos, se llevó a cabo una evaluación objetiva de las predicciones de cada modelo con el conjunto de validación o pruebas. Durante dicha evaluación, se generó un valor de predicción para cada registro del conjunto validación cuyo resultado es cualquiera de las siguientes cuatro posibilidades: verdadero positivo (tp), verdadero negativo (tn), falso positivo (fp) y falso negativo (fn). Las métricas de desempeño que se utilizan en este trabajo son tres: la precisión, la sensibilidad (o *recall*) y el puntaje F1. La precisión consiste en la fracción de predicciones en las que el modelo acierta (ver Ecuación 2). Por otra parte, la fracción del total de registros relevantes que fueron obtenidos se denota como sensibilidad y se calcula en la Ecuación 3. Finalmente, el puntaje F1 representa la media armónica de precisión y sensibilidad. La Ecuación 4 define la métrica de puntaje F1.

$$\text{Precisión} = \frac{tp}{tp + fp}. \tag{2}$$

$$\text{Sensibilidad} = \frac{tp}{tp + fn}. \tag{3}$$

$$\text{Puntaje F1} = 2 \cdot \frac{\text{Precisión} \cdot \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}}. \tag{4}$$

En particular, los valores que se presentan en todas las tablas de este capítulo,

representan una media aritmética de los puntajes F1 calculados de forma separada para cada una de las clases. Esta media se conoce como el puntaje F1 macro-promediado, o macro-F1 por simplicidad.

Finalmente, todos los experimentos se ejecutaron en una estación de trabajo Dell Precision 3420 SFF, con un procesador Intel Xeon E3-1270 v6, memoria RAM de 8GB 2400 MHz DDR4, bajo el sistema operativo Linux Debian 10.6 y utilizando Python 3.6.

A continuación, se presentan los resultados de los modelos generados para cada combinación de las técnicas de balanceo.

IV.1. Resultados para cada configuración de las técnicas de balanceo

Con el objetivo de balancear las clases de los conjuntos de datos, en la Sección III.4 se proponen diversas combinaciones de técnicas de sub-muestreo y sobre-muestreo. Las técnicas de sobre-muestro que se utilizan son SMOTE (O1) y Borderline-SMOTE (O2), mientras que las de submuestreo son All KNN (U1), Cluster-Centroids (U2) y NearMiss (U3). Para cada una de estas combinaciones, se construyeron modelos de clasificación tanto para un enfoque supervisado como semi-supervisado (ver Sección III.5). También, se consideraron las combinaciones de las características extraídas bajo los archivos de configuración `IS09_emotion.conf`, `IS10_paraling.conf`, `IS11_speaker_state.conf` y `Emo_large.conf` de la Sección III.3. Bajo el enfoque supervisado, se utilizan los siguientes algoritmos de aprendizaje: la clasificación de vector de soporte lineal (LSVC), el algoritmo de K vecinos más cercanos (KNN), bosques aleatorios (RF), árboles de decisión (DT), redes neuronales artificiales (NN), naive Bayes (NB), el

algoritmo de impulso adaptativo (AB), el análisis discriminativos lineal (LDA) y el cuadrático (QDA). Asimismo, se utilizan los algoritmos de aprendizaje semi-supervisado LabelPropagation (LP), LabelSpreading (LS) y Self-Training (ST) de la Sección II.3. En particular, el algoritmo ST utiliza los algoritmos LSVC, KNN o RF para la predicción de los registros no etiquetados.

La Tabla XVIII muestra los resultados para los aspectos de emociones, estado mental y paralingüística utilizando el archivo de configuración `IS09_emotion.conf`. El mejor resultado para el aspecto paralingüístico de emociones, considerando únicamente datos etiquetados por consenso y bajo el enfoque supervisado, se obtiene con la combinación de la técnica de sub-muestreo U3 con la de sobre-muestreo O1 (denotado aquí como U3 & O1). Esto bajo el algoritmo de aprendizaje QDA, mientras que para la técnica semi-supervisada se obtiene bajo el algoritmo LP. Por otra parte, el mejor resultado del aspecto paralingüístico de estado mental bajo el enfoque supervisado se obtiene con la combinación U3 & O1 con un resultado de 99.18 %. Bajo el enfoque semi-supervisado, la mejor combinación fue U1 & O1 con un resultado de 97.24 %. Finalmente, bajo el aspecto de paralingüística, se obtiene un resultado del 100 % con las combinaciones U2 & O2, U2 & O1, U3 & O2 y U3 & O1 utilizando el algoritmo de KNN con $K = 3$ (se utilizó este valor en K ya que es el que venía por defecto en la implementación). Del resto de las combinaciones no se obtuvo un resultado debido a que el entrenamiento de los modelos no se completó con éxito. Se conjetura que esto se debe a que los parámetros establecidos en cada una de las técnicas no fue favorable para la distribución de las clases del aspecto de paralingüística. Adicionalmente, cabe mencionar que el proceso de entrenamiento de algunos de los modelos bajo algunas configuraciones, se fue dificultando debido a la dimensionalidad de las características y el alto costo computacional. Esta dificultad también se presentó en otras combinaciones y archivos de configuración que se presentan

en las tablas subsecuentes.

Tabla XVIII. Resultados de la configuración `IS09_emotion.conf` con base a las técnicas supervisadas y semi-supervisadas considerando los aspectos paralingüísticos de emociones, estado mental y paralingüística.

Emociones							
Modelos		U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
supervisado	QDA	76.36 %	77.30 %	77.19 %	79.47 %	83.33 %	84.83 %
semi-supervisado	LP	72.55 %	65.37 %	67.52 %	61.61 %	70.06 %	63.85 %
	LS	71.78 %	65.68 %	66.11 %	60.92 %	70.34 %	66.69 %
Estado Mental							
Modelos		U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
supervisado		QDA=99.16 %	QDA=99.12 %	RF=95.08 %	RF=95.90 %	NB=99.10 %	NB=99.18 %
semi-supervisado	LP	31.24 %	30.54 %	40.16 %	37.70 %	44.26 %	35.25 %
	LS	96.33 %	97.24 %	47.54 %	37.70 %	77.05 %	69.67 %
Paralingüística							
Modelos		U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
supervisado	KNN	-	-	100 %	100 %	100 %	100 %
semi-supervisado	LP	-	-	-	-	-	-
	LS	-	-	-	-	-	-

Por otro lado, la Tabla XIX muestra los resultados para los aspectos paralingüísticos de comportamiento y frases. Respecto al aspecto de comportamiento, el mejor resultado para la técnica supervisada y semi-supervisada se obtuvo en U1 & V, mientras que en el aspecto de frases el mejor resultado se obtuvo en U3 & V. Cabe mencionar que para el aspecto de frases, los registros logran un consenso ya que no existe subjetividad en la respuesta del niño.

La Tabla XX muestra los resultados para los aspectos de emociones, estado mental y paralingüística utilizando el archivo de configuración `IS10_paraling.conf`. El mejor resultado para los aspectos paralingüísticos de emociones y estado mental fue en U1 & O2 bajo el enfoque supervisado, mientras que para la técnica semi-supervisada para emociones fue en U3 & O2 y para estado mental fue en U1 & O1. Nuevamente para el resto de las combinaciones no se obtuvo un resultado debido a que el entrenamiento de los modelos no se completó con éxito.

Tabla XIX. Resultados de la configuración `IS09_emotion.conf` con base a las técnicas supervisadas y semi-supervisadas considerando los aspectos paralingüísticos de comportamiento y frases.

Comportamiento				
Modelos		U1&V	U2&V	U3&V
supervisado		DT=91.11 %	NB=65.78 %	NB=70.91 %
semi-supervisado	LP	14.92 %	28.53 %	34.76 %
	LS	89.83 %	28.53 %	56.65 %
Frases				
Modelos		U1&V	U2&V	U3&V
supervisado		NB=59.58 %	RF=58.90 %	NB=64.38 %
semi-supervisado	LP	32.19 %	31.16 %	32.53 %
	LS	26.71 %	31.85 %	27.74 %

Tabla XX. Resultados de la configuración `IS10_paraling.conf` con base a las técnicas supervisadas y semi-supervisadas considerando los aspectos paralingüísticos de emociones, estado mental y paralingüística.

Emociones							
Modelos		U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
supervisado		NN=81.66 %	NN=81.22 %	RF=76.59 %	NN=79.16 %	RF=78.43 %	NN=76.50 %
semi-supervisado	LP	72.71 %	69.68 %	68.55 %	64.99 %	70.08 %	66.12 %
	LS	74.41 %	71.96 %	67.30 %	64.49 %	74.48 %	72.67 %
Estado Mental							
Modelos		U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
supervisado		NN=99.00 %	NN=98.47 %	RF=97.54 %	RF=95.90 %	RF=89.34 %	NN=92.62 %
semi-supervisado	LP	33.29 %	33.56 %	54.10 %	42.62 %	53.28 %	48.36 %
	LS	94.84 %	97.28 %	64.75 %	49.18 %	66.39 %	60.66 %
Paralingüística							
Modelos		U1&O2	U1&O1	U2&O2	U2&O1	U3&O2	U3&O1
supervisado	KNN	-	-	100 %	100 %	100 %	100 %
semi-supervisado	LP	-	-	-	-	-	-
	LS	-	-	-	-	-	-

Por otro lado, la Tabla XXI muestra los resultados para los aspectos paralingüísticos de comportamiento y frases con el mismo archivo de configuración `IS10_paraling.conf`. Los mejores resultados de comportamiento y frases para la técnica supervisada fue en U1 & V, mientras que para la técnica semi-supervisada de comportamiento fue en U1 & V y para frases fue en U3 & V.

La Tabla XXII muestra los resultados para los aspectos de emociones, estado mental

Tabla XXI. Resultados de la configuración `IS10_paraling.conf` con base a las técnicas supervisadas y semi-supervisadas considerando los aspectos paralingüísticos de comportamiento y frases.

Comportamiento				
Modelos		U1&V	U2&V	U3&V
supervisado		DT=91.91 %	NB=68.14 %	NB=66.48 %
semi-supervisado	LP	13.59 %	28.81 %	32.96 %
	LS	89.66 %	26.18 %	53.74 %
Frases				
Modelos		U1&V	U2&V	U3&V
supervisado		NB=68.15 %	NB=59.24 %	NB=68.83 %
semi-supervisado	LP	39.38 %	37.33 %	38.70 %
	LS	31.51 %	28.42 %	30.82 %

y paralingüística utilizando el archivo de configuración `IS11_speaker_state.conf`. En la tabla de emociones el mejor resultado para la técnica supervisada se obtiene en U3 & O1, mientras que para la técnica semi-supervisada fue en U2 & O2 & V. Respecto al aspecto de estado mental, el mejor resultado con la técnica supervisada se obtiene en U1 & O2 & V, mientras que para la técnica semi-supervisada es en U1 & O2. Similar a los resultados bajo la configuración `IS10_paraling.conf`, no se tienen resultados para el resto de las configuraciones.

La Tabla XXIII muestra los resultados para los aspectos paralingüísticos de comportamiento y frases con el mismo archivo de configuración. Los mejores resultados para la técnica supervisada y semi-supervisada de comportamiento coincidió en ambas bajo la combinación U1 & V, mientras que para el aspecto de frases el mejor resultado de la técnica supervisada fue en U3 & V y para la semi-supervisada en U1 & V.

La Tabla XXIV muestra los resultados para los aspectos paralingüísticos de emociones utilizando el archivo de configuración `Emo_large.conf`. El mejor resultado aquí para la técnica supervisada fue en U1 & O2, mientras que para la técnica semi-supervisada fue en U3 & O2. La Tabla XXV presenta los resultados para los aspectos

Tabla XXII. Resultados de la configuración `IS11_speaker_state.conf` con base a las técnicas supervisadas y semi-supervisadas considerando los aspectos paralingüísticos de emociones, estado mental y paralingüística.

Emociones							
Modelos		U1&O1&V	U2&O2&V	U3&O2&V	U3&O1	U3&O1&V	U1&O1
supervisado		QDA=67.46 %	QDA=66.96 %	KNN=57.68 %	QDA=68.41 %	QDA=56.76 %	-
semi-supervisado	LP	39.83 %	49.56 %	48.97 %	35.11 %	36.31 %	-
	LS	48.93 %	52.61 %	56.61 %	46.15 %	45.95 %	-
Estado Mental							
Modelos		U1&O2	U1&O2&V	U2&O1	U3&O2	U3&O1	U1&O1
supervisado		DT=93.07 %	QDA=93.80 %	RF=81.02 %	QDA=70.42 %	DT=70.71 %	-
semi-supervisado	LP	68.29 %	67.60 %	38.82 %	48.07 %	42.87 %	-
	LS	92.26 %	92.25 %	62.33 %	65.03 %	61.66 %	-
Paralingüística							
Modelos		U1&O2	U1&O2&V	U2&O1	U3&O2	U3&O1	U1&O1
supervisado		-	-	100 %	100 %	100 %	100 %
semi-supervisado	LP	-	-	-	-	-	-
	LS	-	-	-	-	-	-

Tabla XXIII. Resultados de la configuración `IS11_speaker_state.conf` con base a las técnicas supervisadas y semi-supervisadas considerando los aspectos paralingüísticos de comportamiento y frases.

Comportamiento				
Modelos		U1&V	U2&V	U3&V
supervisado		DT=72.17 %	DT=62.74 %	AB=57.97 %
semi-supervisado	LP	44.91 %	35.15 %	37.05 %
	LS	69.02 %	47.47 %	51.81 %
Frases				
Modelos		U1&V	U2&V	U3&V
supervisado		RF=54.79 %	RF=54.45 %	RF=58.21 %
semi-supervisado	LP	29.45 %	27.74 %	23.29 %
	LS	31.85 %	24.66 %	27.05 %

estado mental y paralingüística, donde se puede observar que para el estado mental, los mejores resultados se obtuvieron en U3 & O2.

Finalmente, en la Tabla XXVI se puede observar que para los aspectos paralingüísticos de comportamiento y frases bajo el archivo de configuración `Emo_large.conf`. El mejor resultado para el aspecto de comportamiento bajo las técnicas supervisada y semi-supervisada fue en U3 & V, mientras que para el aspecto de frases fue en U2 & V.

Tabla XXIV. Resultados de la configuración `Emo_large.conf` con base a las técnicas supervisadas y semi-supervisadas considerando los aspectos paralingüísticos de emociones.

Emociones						
Modelos		U1&O2	U1&O1	U2&O2	U3&O2	U3&O2&V
supervisado		QDA=68.45 %	KNN=64.12 %	QDA=67.72 %	KNN=68.36 %	KNN=68.11 %
semi-supervisado	LP	59.38 %	54.88 %	58.92 %	59.82 %	58.94 %
	LS	62.11 %	58.91 %	60.11 %	64.69 %	64.28 %

Tabla XXV. Resultados de la configuración `Emo_large.conf` con base a las técnicas supervisadas y semi-supervisadas considerando los aspectos paralingüísticos de estado mental y paralingüística.

Estado Mental				
Modelos		U2&O2	U2&O1	U3&O2
supervisado	KNN	-	-	81.50 %
semi-supervisado	LP	-	-	52.70 %
	LS	-	-	76.01 %
Paralingüística				
Modelos		U2&O2	U2&O1	U3&O2
supervisado	KNN	100 %	100 %	100 %
semi-supervisado	LP	-	-	-
	LS	-	-	-

Tabla XXVI. Resultados de la configuración `Emo_large.conf` con base a las técnicas supervisadas y semi-supervisadas considerando los aspectos paralingüísticos de comportamiento y frases.

Comportamiento			
Modelos		U2&V	U3&V
supervisado	DT	61 %	63.80 %
semi-supervisado	LP	34.43 %	35.53 %
	LS	48.52 %	54.85 %
Frases			
Modelos		U2&V	U3&V
supervisado	RF	56 %	55 %
semi-supervisado	LP	31.85 %	27.40 %
	LS	33.90 %	33.80 %

Se concluye que los resultados obtenidos en cada unos de los aspectos paralingüísticos fue con los algoritmos supervisados, esto debido a que el aprendizaje se basa en el

corpus de datos previamente etiquetado. En particular, el mejor resultado para el aspecto paralingüístico de emociones fue de 84.83 %, mientras que para el aspecto de estado mental fue de 99.18 %, ambos resultados se obtuvieron con el archivo de configuración `IS09_emotion.conf`. Respecto al aspecto paralingüístico de comportamiento, se obtuvo un puntaje de 91.91 % y para el de frases 68.83 %, en ambos casos se uso el archivo de configuración de `IS10_paraling.conf`. Por último, el mejor resultado del aspecto de paralingüística fue de 100 % con todos los archivos de configuración: `IS09_emotion.conf`, `IS10_paraling.conf`, `IS11_speaker_state.conf` y `Emo_large.conf`. Este resultado se debe a la ausencia de subjetividad en el análisis.

IV.2. Modelos independientes del hablante

En esta sección se muestran los resultados de los modelos generados bajo un enfoque independiente del hablante, descrito en la Sección III.6.

Las tablas XXVII-XXIX muestran los resultados de los algoritmos LSVC, KNN y RF, respectivamente, para los aspectos paralingüísticos de emociones, comportamiento, estado mental, frases y paralingüística. Se hace la comparación de estos modelos supervisados versus los semi-supervisados (ST, LP y LS) mostrando en rojo los mejores resultados. La Tabla XXVII muestra que los mejores resultados para el modelo supervisado y semi-supervisado utilizando LSVC fue con el archivo de configuración `IS10_paraling.conf`. De aquí se observa que solamente para los aspectos de estado mental y paralingüística la técnica semi-supervisada fue mejor que la supervisada. Esto probablemente se debe a la calidad de las estimaciones de las etiquetas en los registros con un bajo nivel de confiabilidad. Por otro lado, la Tabla XXVIII (correspondiente al KNN) en general tiene los mejores resultados en los archivos de configuración

`IS10_paraling.conf` y `IS09_emotion.conf`. A diferencia de los resultados de LSVC, los mejores resultados se obtienen con KNN con $K = 3$. Por otra parte, respecto a los resultados de RF de la Tabla XXIX, los resultados fueron muy diversos en referencia a los archivos de configuración. En este caso, las técnicas semi-supervisadas obtienen mejores resultados solamente para los aspectos de emociones y paralingüística. Finalmente, cabe destacar que el algoritmo semi-supervisado ST obtuvo los mejores resultados para casi todos los casos.

Tabla XXVII. Resultados del algoritmo LSVC con base al modelo independiente del hablante considerando los aspectos paralingüísticos de emociones, comportamiento,estado mental, frases y paralingüística.

Emociones (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
LSVC	11.00 %	25.90 %	8.30 %	12.90 %
ST	12.80 %	24.50 %	15.60 %	14.80 %
LP	15.00 %	18.00 %	15.70 %	15.40 %
LS	18.90 %	21.50 %	20.80 %	20.00 %
Comportamiento (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
LSVC	40.90 %	46.80 %	10.00 %	26.60 %
ST	43.30 %	46.00 %	18.00 %	11.30 %
LP	28.80 %	30.10 %	26.40 %	26.50 %
LS	38.80 %	38.90 %	32.60 %	34.50 %
Estado Mental (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
LSVC	34.30 %	36.20 %	32.10 %	32.90 %
ST	34.80 %	47.60 %	25.40 %	37.10 %
LP	25.00 %	28.90 %	30.20 %	31.90 %
LS	32.80 %	36.30 %	38.30 %	35.60 %
Frases (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
LSVC	49.50 %	70.00 %	14.80 %	38.80 %
ST	35.00 %	65.60 %	12.90 %	31.60 %
LP	22.70 %	23.20 %	21.20 %	21.90 %
LS	20.40 %	28.70 %	27.40 %	23.90 %
Paralingüística (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
LSVC	29.40 %	44.80 %	12.40 %	30.80 %
ST	39.30 %	54.10 %	6.00 %	26.30 %
LP	19.70 %	16.40 %	19.80 %	17.30 %
LS	40.60 %	28.20 %	25.70 %	30.60 %

La Tabla XXX concentra los mejores resultados de las tablas XXVII-XXIX

Tabla XXVIII. Resultados del algoritmo KNN con base al modelo independiente del usuario considerando los aspectos paralingüísticos de emociones, comportamiento,estado mental, frases y paralingüística.

Emociones (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
KNN	21.00 %	23.00 %	19.50 %	20.70 %
ST	19.90 %	22.60 %	20.00 %	19.00 %
LP	14.60 %	18.50 %	15.40 %	14.40 %
LS	19.20 %	21.00 %	20.40 %	19.00 %
Comportamiento (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
KNN	40.80 %	40.30 %	33.40 %	35.10 %
ST	40.50 %	38.80 %	33.20	36.10 %
LP	28.80 %	30.10 %	26.40 %	26.50 %
LS	38.80 %	38.90 %	32.60 %	34.50 %
Estado Mental (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
KNN	33.10 %	41.10 %	36.10 %	37.90 %
ST	33.20 %	38.80 %	35.00 %	36.50 %
LP	25.00 %	28.90 %	30.20 %	31.90 %
LS	32.80 %	36.30 %	38.30 %	35.60 %
Frases (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
KNN	30.60 %	36.20 %	30.40 %	29.70 %
ST	22.80 %	27.30 %	28.70 %	23.10 %
LP	20.10 %	23.20 %	20.70 %	21.90 %
LS	22.00 %	28.70 %	26.60 %	20.40 %
Paralingüística (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
KNN	42.20 %	28.70 %	22.40 %	29.60 %
ST	42.00 %	31.00 %	25.20 %	33.90 %
LP	20.60 %	17.30 %	14.70 %	15.50 %
LS	40.80 %	27.30 %	23.10 %	29.90 %

comparando las técnicas supervisadas (LSVC, KNN y RF) y las semi-supervisadas (ST, LP y LS) con los resultados reportados en (Pérez-Espinosa *et al.*, 2018a) y (Pérez-Espinosa *et al.*, 2018b), los cuales utilizan un enfoque meramente supervisado. Se puede observar que los resultados aquí obtenidos no superaron a los reportados en (Pérez-Espinosa *et al.*, 2018b), que considera únicamente los aspectos de emociones y comportamiento. Por otro lado, los resultados sí fueron superiores a los reportados en (Pérez-Espinosa *et al.*, 2018a) para los aspectos de estado mental y paralingüística.

Analizando la situación de estos resultados bajo el modelo independiente del

Tabla XXIX. Resultados del algoritmo RF con base al modelo independiente del usuario considerando los aspectos paralingüísticos de emociones, comportamiento, estado mental, frases y paralingüística.

Emociones (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
RF	30.60 %	34.30 %	31.70 %	33.60 %
ST	29.50 %	34.60 %	30.80 %	34.10 %
LP	15.00 %	18.80 %	16.20 %	14.30 %
LS	18.90 %	21.20 %	19.70 %	19.20 %
Comportamiento (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
RF	49.90 %	46.60 %	47.80 %	47.80 %
ST	48.00 %	47.40 %	45.40 %	47.70 %
LP	28.80 %	30.10 %	26.40 %	26.50 %
LS	38.80 %	38.90 %	32.60 %	34.50 %
Estado Mental (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
RF	46.50 %	42.70 %	47.50 %	43.10 %
ST	45.00 %	42.50 %	46.20 %	44.50 %
LP	25.00 %	28.90 %	30.20 %	31.90 %
LS	32.80 %	36.30 %	38.30 %	35.60 %
Frases (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
RF	54.60 %	52.20 %	54.70 %	62.30 %
ST	54.50 %	59.20 %	48.80 %	50.70 %
LP	23.40 %	23.20 %	19.80 %	21.90 %
LS	21.10 %	28.70 %	25.10 %	20.40 %
Paralingüística (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
RF	50.70 %	51.10 %	53.30 %	56.20 %
ST	59.50 %	61.10 %	63.20 %	60.00 %
LP	23.60 %	16.60 %	16.60 %	16.30 %
LS	37.80 %	31.00 %	23.70 %	24.60 %

hablante, se procedió a calcular el promedio de los niveles de confiabilidad para cada uno de los usuarios del conjunto de datos. En particular, se analizó el aspecto de emociones considerando los siguientes tres usuarios: Ramón con un promedio de 0.863095, Ada con 0.879121 y Ángel con 0.948052. En la Tabla XXXI, se muestran las diversas incidencias de estos usuarios dentro de las ocho clases del aspecto de emociones. Se observa allí que los usuarios no necesariamente aparecen en las ocho clases de la tabla, y que el promedio se calcula con base a los registros que tiene cada clase (i.e., no necesariamente es el mejor promedio de la clase). Así, es probable que el entrenamiento de los modelos

Tabla XXX. Resultados finales de los algoritmos LSVC, KNN y RF del modelo independiente del hablante.

LSVC					
Modelos	Emociones	Comportamiento	Estado Mental	Frases	Paralingüística
supervisado	25.90 %	46.80 %	36.20 %	70.00 %	44.80 %
semi–supervisado	24.50 %	46.00 %	47.60 %	65.60 %	54.10 %
(Pérez-Espinosa *et al.*, 2018a)	23.20 %	30.80 %	31.70 %	81.40 %	28.50 %
(Pérez-Espinosa *et al.*, 2018b)	56.33 %	65.55 %	-	-	-
KNN					
Modelos	Emociones	Comportamiento	Estado Mental	Frases	Paralingüística
supervisado	23.00 %	40.80 %	41.10 %	36.20 %	42.20 %
semi–supervisado	22.60 %	40.50 %	38.80 %	28.70 %	42.00 %
(Pérez-Espinosa *et al.*, 2018a)	23.20 %	30.80 %	31.70 %	81.40 %	28.50 %
(Pérez-Espinosa *et al.*, 2018b)	56.33 %	65.55 %	-	-	-
RF					
Modelos	Emociones	Comportamiento	Estado Mental	Frases	Paralingüística
supervisado	34.30 %	49.90 %	47.50 %	62.30 %	56.20 %
semi–supervisado	34.60 %	48.00 %	46.20 %	59.20 %	63.20 %
(Pérez-Espinosa *et al.*, 2018a)	23.20 %	30.80 %	31.70 %	81.40 %	28.50 %
(Pérez-Espinosa *et al.*, 2018b)	56.33 %	65.55 %	-	-	-

considerando usuarios distintos a los de pruebas y/o validación tenga un sesgo en la predicción de ciertas clases para cada aspecto paralingüístico, lo cual impacta en la calidad de los resultados.

Tabla XXXI. Incidencia de los usuarios con el mayor nivel de confiabilidad promedio dentro de las clases del aspecto paralingüístico de emociones y bajo el modelo independiente del hablante.

Clases	**Ramón**	**Ada**	**Ángel**
Ninguno	9	2	3
Tristeza	5	1	-
Neutral	5	9	7
Desprecio	4	1	-
Enojo	3	-	-
Miedo	2	-	-
Felicidad	1	-	1

IV.3. Modelos dependientes del hablante

En esta sección se muestran los resultados de los modelos generados bajo un enfoque dependiente del hablante, descrito en la Sección III.6. Bajo este modelo, aunque los registros tanto del conjunto entrenamiento como el de prueba son distintos, algunos usuarios que aparecen en el conjunto de entrenamiento también lo hacen en el de pruebas.

Las tablas XXXIII-XXXIV muestran los resultados de los algoritmos LSVC, KNN yRF, respectivamente, para los aspectos paralingüísticos de emociones, comportamiento, estado mental, frases y paralingüística bajo el enfoque dependiente del hablante. Similar a lo explicado en la Sección IV.2, se hace una comparación de estos modelos supervisados con los semi-supervisados (ST, LP y LS). En la Tabla XXXII se observa que los mejores resultados tanto para los modelos supervisados (LSVC) como semi-supervisados (ST, LP y LS) se obtienen con el archivo de configuración `IS10_paraling.conf` para todas las configuraciones. Esto difiere de las otras dos tablas, donde los mejores resultados se encuentran dispersos en los diferentes archivos de configuración. Cabe enfatizar que bajo este modelo dependiente del hablante, la técnica semi-supervisada fue mejor que la supervisada para los aspectos de estado mental, frases y paralingüística para las tres tablas (excepto en el aspecto paralingüístico de frases de la Tabla XXXIII). Es importante observar también que, a diferencia del modelo independiente del hablante, la calidad de los resultados del modelo dependiente son superiores en general.

Por último, la Tabla XXXV muestra un concentrado de los mejores resultados de las tablas XXXII-XXXIV, comparando las técnicas supervisadas (LSVC, KNN y RF) y las semi-supervisadas (ST, LP y LS) con los resultados reportados en (Pérez-Espinosa *et al.*, 2018a) y (Pérez-Espinosa *et al.*, 2018b). Se puede observar que, los resultados

Tabla XXXII. Resultados del algoritmo LSVC con base al modelo dependiente del usuario considerando los aspectos paralingüísticos de emociones, comportamiento, estado mental, frases y paralingüística.

Emociones (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
LSVC	69.50 %	84.60 %	9.40 %	54.40 %
ST	50.50 %	83.40 %	7.10 %	37.50 %
LP	65.50 %	69.60 %	57.50 %	68.50 %
LS	54.40 %	64.10 %	53.30 %	64.00 %
Comportamiento (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
LSVC	52.90 %	64.90 %	26.50 %	47.40 %
ST	53.00 %	60.90 %	20.30 %	47.70 %
LP	23.90 %	26.20 %	21.60 %	21.70 %
LS	55.20 %	60.20 %	49.20 %	47.40 %
Estado Mental (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
LSVC	88.70 %	92.70 %	37.40 %	81.60 %
ST	83.70 %	95.10 %	37.60 %	57.20 %
LP	92.00 %	94.10 %	79.30 %	96.10 %
LS	77.00 %	79.80 %	68.70 %	75.90 %
Frases (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
LSVC	41.90 %	59.90 %	8.30 %	34.90 %
ST	44.40 %	60.60 %	10.10 %	29.50 %
LP	18.60 %	15.00 %	18.70 %	15.40 %
LS	18.90 %	16.40 %	19.90 %	16.50 %
Paralingüística (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
LSVC	92.60 %	94.80 %	14.60 %	80.40 %
ST	83.40 %	95.00 %	19.30 %	66.00 %
LP	77.80 %	85.90 %	57.90 %	83.90 %
LS	69.30 %	74.00 %	49.60 %	73.70 %

bajo el enfoque dependiente del hablante, son competitivos y superan a estos dos trabajos previos en la mayoría de los aspectos paralingüísticos. También se observa que en general se obtienen muy buenos resultados bajo la técnica semi-supervisada, principalmente en los aspectos de estado mental, frases y paralingüística.

Tabla XXXIII. Resultados del algoritmo KNN con base al modelo dependiente del usuario considerando los aspectos de emociones, comportamiento, estado mental, frases y paralingüística.

Emociones (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
KNN	66.20 %	72.80 %	61.00 %	67.30 %
ST	66.20 %	71.90 %	58.40 %	65.90 %
LP	67.00 %	70.10 %	55.70 %	68.20 %
LS	55.80 %	63.90 %	52.10 %	61.50 %
Comportamiento (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
KNN	58.60 %	58.30 %	47.70 %	47.90 %
ST	55.10 %	55.20 %	42.80 %	44.30 %
LP	24.70 %	26.10 %	20.20 %	22.10 %
LS	55.30 %	58.50 %	48.40 %	48.00 %
Estado Mental (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
KNN	74.30 %	73.40 %	72.40 %	74.90 %
ST	74.30 %	73.30 %	70.20 %	76.40 %
LP	91.70 %	96.40 %	85.40 %	95.80 %
LS	77.30 %	80.50 %	68.80 %	75.00 %
Frases (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
KNN	23.40 %	25.50 %	22.90 %	18.80 %
ST	19.90 %	17.70 %	18.60 %	18.10 %
LP	18.60 %	15.00 %	18.70 %	15.40 %
LS	18.90 %	16.40 %	19.90 %	16.50 %
Paralingüística (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
KNN	75.90 %	84.10 %	62.80 %	76.50 %
ST	75.70 %	81.60 %	61.90 %	77.20 %
LP	76.20 %	87.10 %	58.00 %	79.60 %
LS	66.20 %	76.50 %	49.60 %	69.70 %

Tabla XXXIV. Resultados del algoritmo RF con base al modelo dependiente del usuario considerando los aspectos de emociones, comportamiento, estado mental, frases y paralingüística.

Emociones (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
RF	82.90 %	85.70 %	83.50 %	86.50 %
ST	81.30 %	81.10 %	82.30 %	82.40 %
LP	64.90 %	70.30 %	57.00 %	66.80 %
LS	54.60 %	64.00 %	54.20 %	61.30 %
Comportamiento (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
RF	67.40 %	63.10 %	66.80 %	68.80 %
ST	62.00 %	62.40 %	64.50 %	62.60 %
LP	26.10 %	25.40 %	20.30 %	21.00 %
LS	56.70 %	58.70 %	49.30 %	47.70 %
Estado Mental (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
RF	96.00 %	97.70 %	97.10 %	90.30 %
ST	96.50 %	98.50 %	97.40 %	95.60 %
LP	90.60 %	97.30 %	82.00 %	95.50 %
LS	76.30 %	80.30 %	69.20 %	75.60 %
Frases (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
RF	45.20 %	48.60	46.60 %	53.40 %
ST	47.00 %	40.90 %	48.10 %	47.90 %
LP	18.60 %	15.00 %	18.70 %	15.40 %
LS	18.90 %	16.44 %	19.90 %	16.50 %
Paralingüística (Puntaje F1)				
configuración	IS09	IS10	IS11	EMO-LARGE
RF	92.30 %	93.50 %	93.20 %	92.40 %
ST	90.10 %	90.70 %	90.20 %	90.30 %
LP	80.10 %	87.00 %	56.50 %	83.50 %
LS	68.60 %	73.40 %	54.10 %	68.80 %

Tabla XXXV. Resultados finales de los algoritmos LSVC, KNN y RF del modelo dependiente del hablante.

LSVC					
Modelos	Emociones	Comportamiento	Estado Mental	Frases	Paralingüística
supervisado	84.60 %	64.90 %	92.70 %	59.90 %	94.80 %
semi–supervisado	83.40 %	60.90 %	96.10 %	60.60 %	95.00 %
(Pérez-Espinosa *et al.* 2018a)	23.20 %	30.80 %	31.70 %	81.40 %	28.50 %
(Pérez-Espinosa *et al.* 2018b)	56.33 %	65.55 %	-	-	-
KNN					
Modelos	Emociones	Comportamiento	Estado Mental	Frases	Paralingüística
supervisado	72.80 %	58.60 %	74.90 %	25.50 %	84.10 %
semi–supervisado	71.90 %	58.50 %	96.40 %	19.90 %	87.10 %
(Pérez-Espinosa *et al.* 2018a)	23.20 %	30.80 %	31.70 %	81.40 %	28.50 %
(Pérez-Espinosa *et al.* 2018b)	56.33 %	65.55 %	-	-	-
RF					
Modelos	Emociones	Comportamiento	Estado Mental	Frases	Paralingüística
supervisado	86.50 %	68.80 %	97.70 %	53.40 %	93.50 %
semi–supervisado	82.40 %	62.60 %	98.50 %	48.10 %	90.70 %
(Pérez-Espinosa *et al.* 2018a)	23.20 %	30.80 %	31.70 %	81.40 %	28.50 %
(Pérez-Espinosa *et al.* 2018b)	56.33 %	65.55 %	-	-	-

Capítulo V

Diseño del prototipo

En este capítulo se presenta el diseño de un prototipo de un videojuego de interacción por voz para niños entre 6 y 12 años. El diseño propuesto contempla que el videojuego reconozca un conjunto de comandos de voz, junto con un aspecto paralingüístico de emociones como datos de entrada. El videojuego detecta las emociones de los usuarios utilizando el modelo de aprendizaje para la detección de emociones de la Sección III.5. Para reconocer los comandos de voz, fue necesario generar un nuevo corpus de audios de niños que permitieran entrenar un nuevo modelo de aprendizaje para tal fin.

A continuación, se describe el proceso de construcción del nuevo corpus.

V.1. Corpus de comandos

Se recopilaron audios de 22 niños entre los 6 y 12 años de edad. Los audios contienen seis comandos específicos o etiquetas: 'izquierda', 'derecha', 'sube', 'baja', 'vuela' e 'inicia'. De cada niño se obtuvieron 30 segmentos de audios (cinco por etiqueta), por lo que se recabaron un total de 660 audios en total.

La obtención de los segmentos de audio de los niños se hizo a través de la aplicación móvil WhatsApp, con ayuda y autorización de los padres. Dentro de las indicaciones que se les dio a los padres, la más importante fue el de eliminar los sonidos de fondo (televisión, electrodomésticos, mascotas o voces) y que la grabación incluyera palabras completas. De los 660 audios recolectados, nueve no cumplieron con estas indicaciones y fueron descartados. La Tabla XXXVI muestra la distribución de los 651 audios del

corpus de comandos para el videojuego.

Tabla XXXVI. Distribución de los 651 audios del corpus de comandos para el videojuego.

Etiquetas	**# Audios**
sube	110
izquierda	109
derecha	109
inicia	108
vuela	108
baja	107

V.2. Modelos de clasificación de comandos

Se realizó la extracción de características de los 651 audios de comandos utilizando openSmile con los archivos de configuración `IS09_emotion.conf` y `IS10_paraling.conf`, descritos en la Sección III.3. Para generar los modelos de clasificación de los comandos, se utilizaron los algoritmos de aprendizaje supervisado: LinearSVC, KNN y RF (descritos en la Sección III.5.3). Los modelos de clasificación se generaron considerando el modelo independiente del hablante, tal como se describe en la Sección III.6. Asimismo, se utilizó un 70 % del corpus para el conjunto entrenamiento y un 30 % para el conjunto de validación. Se empleó esta técnica de muestreo debido a que el conjunto de datos tiene pocos registros, donde comúnmente se recomienda utilizar del 50-70 % del corpus para el entrenamiento (Xu y Goodacre, 2018).

La Tabla XXXVII muestra el desempeño de los modelos de clasificación entrenados para cada archivo de configuración de extracción de características. Allí se muestra que los mejores resultados se obtienen con la configuración `IS10_paraling.conf` y, en especial, con el modelo generado por RandomForest. Este modelo fue el que se utilizó dentro del prototipo.

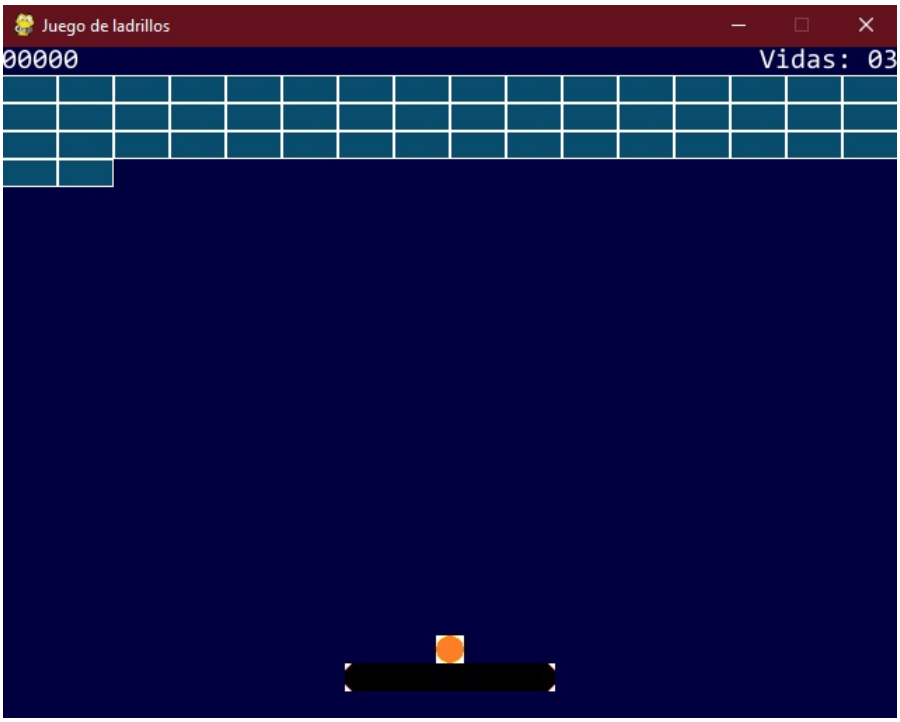
Tabla XXXVII. Desempeño de los modelos de clasificación para el corpus de comandos.

Archivo de configuración	LinearSVC	KNN	RF
IS09_emotion.conf	81.70 %	35.00 %	91.90 %
IS10_paraling.conf	85.60 %	38.60 %	93.30 %

V.3. El videojuego

El videojuego trata de un conjunto de bloques o "ladrillos" que deben romperse por una pelota que rebota a través de una barra movible horizontalmente por el usuario. Este videojuego, conocido como *Breakout* o *Brick breaker*, se desarrolló en 1976 por Nolan Bushnell y Steve Bristow y, a la fecha, cuenta con cientos de versiones para una gran variedad de plataformas. La Figura 9 presenta una pantalla de ejemplo de este videojuego.

Figura 9. Pantalla de ejemplo del videojuego *Breakout*.



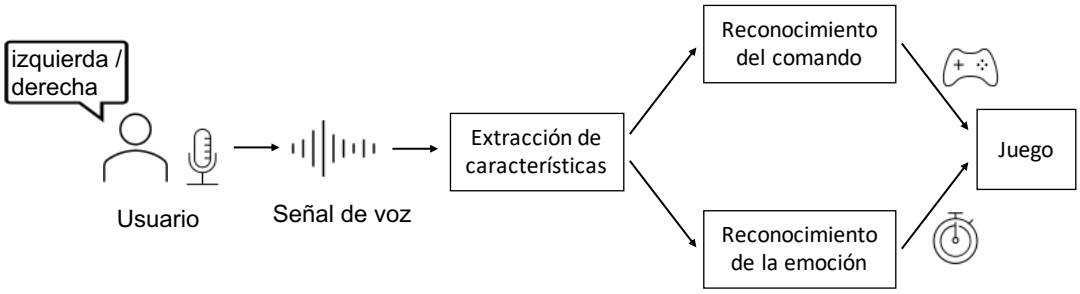
La decisión de utilizar este videojuego radica principalmente en su simplicidad,

lo que facilitó su desarrollo y la incorporación de la interacción por voz a través de los modelos de clasificación. Tanto el videojuego como el módulo de interacción por voz se desarrollaron en Python 3.7. En particular, se utilizaron las librerías Pygame, PyAudio y Wave. Pygame se utilizó para implementar la interfaz del juego, junto con las animaciones y el manejo del teclado. La librería PyAudio permitió realizar la transmisión en tiempo real del audio, capturando las ordenes del usuario. Posterior a ello, con la librería Wave se transcribe el audio a un archivo con formato WAV. A este audio, se le realiza la extracción de características utilizando openSmile con el archivo de configuración `IS10_paraling`. Después, se aplica el algoritmo RF para clasificar el audio por algunos de los comandos permitidos en el juego.

La Figura 10 muestra el diagrama de bloques del prototipo, el cual consiste de cuatro pasos principales. Primero, el usuario dicta un comando de voz que se almacena en un archivo de audio digital con formato WAV. Este paso incluye un proceso de segmentación de audio, que almacena exclusivamente una palabra a la vez por archivo. Posteriormente, se envía el archivo WAV a un módulo de extracción de características acústicas de openSmile utilizando la configuración `IS10_paraling.conf`. El resultado se procesa utilizando dos modelos de clasificación: uno diseñado para el reconocimiento del comando de voz (ver Sección V.2) y el otro especializado para el reconocimiento de las emociones (ver Sección III.6). Finalmente, se envían los resultados de los modelos de clasificación al videojuego. El comando de voz tiene un efecto directo en el movimiento horizontal de la barra dentro del juego, donde solamente requiere de las instrucciones ‘izquierda’ y ‘derecha’ omitiendo el resto de los comandos (‘inicia’, ‘sube’, ‘baja’ y ‘vuela’). Por otro lado, las emociones detectadas influyen en la velocidad del movimiento de la pelota, impactando directamente en la dificultad del juego.

A continuación, se describe la forma como se tratan las emociones en los niveles de

Figura 10. Diagrama de bloques del prototipo.



dificultad del juego.

V.4. Inducción de emociones en el videojuego

En este trabajo, se parte de la premisa de que diferentes niveles de dificultad del juego corresponden a diferentes emociones del usuario. Esta premisa se basa en los estudios de Chanel *et al.* (2011) y (Liu *et al.*, 2009). Para definir la interacción de las emociones con los niveles de dificultad del juego, se propone un diagrama con tres estados: el Nivel I (que representa un juego fácil y lento), el Nivel II (que trata de un juego con un nivel de dificultad moderada) y el Nivel III (que denota un juego rápido y difícil). Pasar de un nivel de dificultad a otro en el juego representa una transición de estado, el cuál se origina a partir de las emociones detectadas en la voz del usuario.

El modelo de clasificación de emociones descrito en la Sección III.6, está diseñado para ocho tipo de emociones: desprecio, enojo, felicidad, miedo, neutral, sorpresa, tristeza y ninguno. Para simplificar la transición de estados, se aglutinan las ocho emociones en tres grupos que aquí se denominan: baja, media y alta. El grupo baja se compone de las emociones de desprecio, enojo y tristeza. El grupo alta trata de las emociones de sorpresa y felicidad. Finalmente, el grupo media trata de una emoción neutral o de la imposibilidad de detectar alguna emoción. Este agrupamiento se basa

en el modelo circunflejo de Russell (1980), el cual se muestra en la Figura 11.

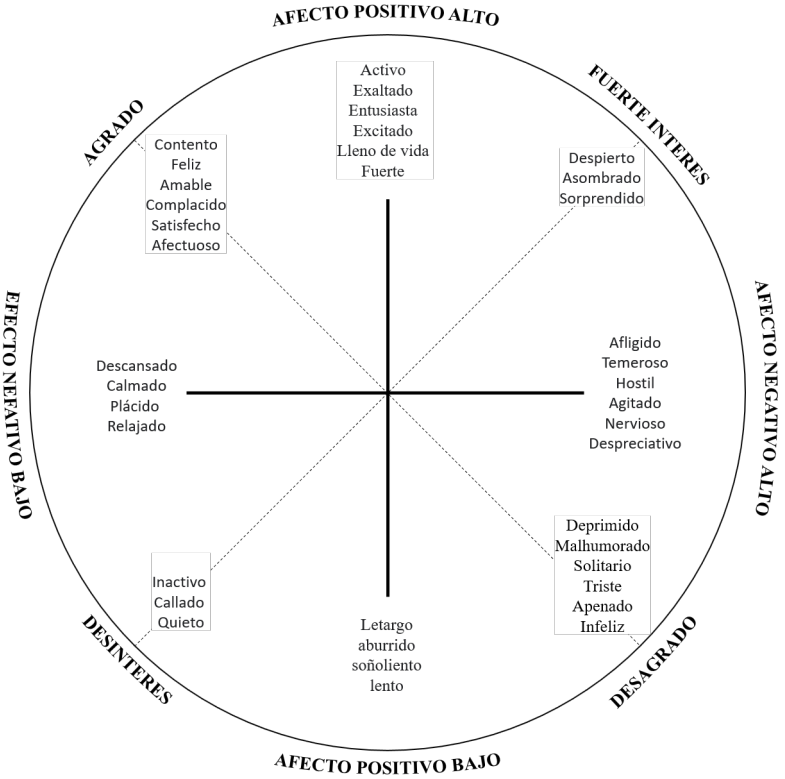


Figura 11. Ejemplos de los estados emocionales del modelo circunflejo de Russell (1980).

La Figura 12 muestra el diagrama de flujo de estados para el ajuste dinámico del nivel de dificultad del videojuego propuesto. En este diagrama, el juego inicia en el Nivel I, y se mantiene allí mientras el usuario tenga una emoción distinta a la del grupo alta. Si durante el juego el usuario manifiesta una emoción del grupo alta, entonces el juego incrementa su velocidad pasando de un juego fácil a un juego moderado (Nivel II). Finalmente, si el usuario sigue manifestado este tipo de emociones, pasará del Nivel II al III. Estas transiciones se proponen con base a la premisa de que los usuarios con un alto afecto positivo, tendrían un mayor compromiso (o *engagement*) en el juego que requiere de un mayor reto o dificultad del mísmo.

Por otro lado, si el usuario en el Nivel III manifiesta en su voz una emoción del

grupo baja, entonces pasará a un Nivel II del juego. Esto mismo ocurrirá del Nivel II al I para este tipo de emociones. La premisa para este las transiciones en el descenso de niveles, es que el usuario con sentimientos negativos expresa aburrimiento o falta de *engagement* hacia el juego. Así, se pretende disminuir los estímulos negativos y el estrés del usuario facilitando las condiciones del juego.

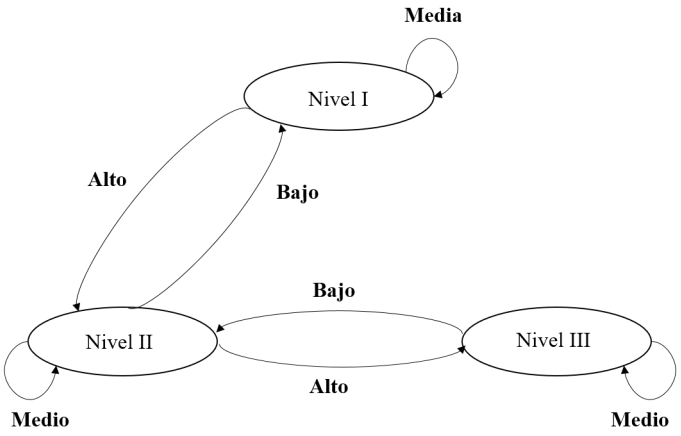


Figura 12. Diagrama de flujo de estados del prototipo.

Capítulo VI

Conclusiones y trabajo futuro

Hoy en día, la computación paralingüística comprende una gran variedad de métodos, herramientas y técnicas que permiten el reconocimiento de afecto, emociones y rasgos de personalidad que se expresan a través del habla y del lenguaje. Bajo este marco computacional, en el presente trabajo se desarrollaron modelos de clasificación acústica de niños para el perfilado de usuarios en sistemas de interacción basados en voz. En particular, se utilizó el corpus de Pérez-Espinosa *et al.* (2018b) que contiene audios de niños hispanohablantes entre seis y doce años de edad, comprendiendo cinco aspectos paralingüísticos etiquetados con diferentes niveles de confiabilidad.

Los modelos de clasificación acústica se construyeron empleando técnicas de aprendizaje supervisado y semi-supervisado. Los resultados experimentales muestran que, en general, los modelos generados a través de técnicas de aprendizaje semi-supervisado tienen un mejor desempeño que los generados por técnicas de aprendizaje supervisado. Esto se debe a la reducción de la subjetividad del etiquetado. En especial, se obtuvieron los mejores resultados utilizando el algoritmo de Random Forest con el conjunto de características acústicas extraidas con la configuración `IS09_emotion.conf`.

Por otra parte, se propuso el diseño de un prototipo de un videojuego de interacción por voz. El prototipo trata el aspecto paralingüístico de emociones del usuario para determinar el nivel de dificultad del videojuego. Aunque se completó el diseño y se implementó el prototipo, queda como trabajo futuro el aplicar esta metodología de

diseño para el desarrollo de videojuegos que involucre aspectos paralingüísticos.

Por último, como trabajo futuro adicional, se propone implementar los modelos de clasificación acústica en otras aplicaciones que involucren otros aspectos de emociones, estado mental y comportamiento. Se conjetura que estas aplicaciones serían útiles para el desarrollo de videojuegos serios.

Referencias bibliográficas

Abercrombie, D. (1968). Paralanguage. *International Journal of Language & Communication Disorders*, **3**(1): 55–59.

Andreas, Pellom, y Cole (2003). Children's speech recognition with application to interactive books and tutors. En *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pp. 186–191, Nov.

Cestero Mancera, A. M. (2006). La comunicación no verbal y el estudio de su incidencia en fenómenos discursivos como la ironía.

Chanel, G., Rebetez, C., Bétrancourt, M., y Pun, T. (2011). Emotion assessment from physiological signals for adaptation of game difficulty. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, **41**(6): 1052–1063.

Chapelle, O., Schölkopf, B., Zien, A., *et al.* (2006). Semi-supervised learning, vol. 2. *Cambridge: MIT Press. Cortes, C., & Mohri, M.(2014). Domain adaptation and sample bias correction theory and algorithm for regression. Theoretical Computer Science*, **519**: 103126.

Chawla, N. V., Bowyer, K. W., Hall, L. O., y Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, **16**: 321–357.

Cucchiarini, C. y Van hamme, H. (2013). *The JASMIN Speech Corpus: Recordings of Children, Non-natives and Elderly People*, pp. 43–59. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-642-30910-6.

Dhupati, L. S., Kar, S., Rajaguru, A., y Routray, A. (2010). A novel drowsiness detection scheme based on speech analysis with validation using simultaneous eeg recordings. En *2010 IEEE international conference on automation science and engineering*, pp. 917–921. IEEE.

Douzas, G., Bacao, F., y Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and smote. *Information Sciences*, **465**: 1–20.

Eyben, F., Wöllmer, M., y Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. En *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, p. 1459–1462, New York, NY, USA. Association for Computing Machinery.

Han, H., Wang, W.-Y., y Mao, B.-H. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning. En D.-S. Huang, X.-P. Zhang, y G.-B. Huang (eds.), *Advances in Intelligent Computing*, pp. 878–887, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hollien, H., DeJong, G., Martin, C. A., Schwartz, R., y Liljegren, K. (2001). Effects of ethanol intoxication on speech suprasegmentals. *The Journal of the Acoustical Society of America*, **110**(6): 3198–3206.

Ishi, C. T., Ishiguro, H., y Hagita, N. (2008). Automatic extraction of paralinguistic information using prosodic features related to f0, duration and voice quality. *Speech Communication*, **50**(6): 531 – 543.

Liu, C., Agrawal, P., Sarkar, N., y Chen, S. (2009). Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback. *International Journal of Human-Computer Interaction*, **25**(6): 506–529.

Mani, I. y Zhang, I. (2003). knn approach to unbalanced data distributions: a case study involving information extraction. En *Proceedings of workshop on learning from imbalanced datasets*, Vol. 126.

Martínez-Miranda, J., Pérez-Espinosa, H., Espinosa-Curiel, I., Avila-George, H., y Rodríguez-Jacobo, J. (2018). Age-based differences in preferences and affective reactions towards a robot's personality during interaction. *Computers in Human Behavior*, **84**: 245 – 257.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., y Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**: 2825–2830.

Pérez-Espinosa, H., Avila-George, H., Martınez-Miranda, J., Espinosa-Curiel, I., Rodriguez-Jacobo, J., y Cruz-Mendoza, H. A. (2018a). Children age and gender classification based on speech using convnets. *Research in Computing Science*, **147**(4): 23–35.

Pérez-Espinosa, H., Martínez-Miranda, J., Avila-George, H., y Espinosa-Curiel, I. (2018b). Analyzing children's affective reactions and preferences towards social robots using paralinguistic and self-reported information. *Journal of Intelligent & Fuzzy Systems*, (Preprint): 1–12.

Pérez-Espinosa, H., Martínez-Miranda, J., Espinosa-Curiel, I., Rodríguez-Jacobo, J., Villaseñor-Pineda, L., y Avila-George, H. (2020). Iesc-child: An interactive emotional children's speech corpus. *Computer Speech & Language*, **59**: 55–74.

Poyatos, F. (1993). *Paralanguage: A linguistic and interdisciplinary approach to interactive speech and sounds*, Vol. 92. John Benjamins Publishing.

Quilis, A. y Fernández, J. A. (1969). Curso de fonética y fonología españolas: para estudiantes angloamericanos.

Rafii, Z. y Pardo, B. (2012). Music/voice separation using the similarity matrix. En *ISMIR*, pp. 583–588.

Rizos, G. y Schuller, B. W. (2020). Average jane, where art thou?–recent avenues in efficient machine learning under subjectivity uncertainty. En *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 42–55. Springer.

Rodríguez, L. B. (2007). Aproximación al paralenguaje. *Hesperia: Anuario de filología hispánica*, (10): 83–97.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, **39**(6): 1161.

Safavi, S., Russell, M., y Jančovič, P. (2014). Identification of age-group from children's speech by computers and humans. En *Fifteenth Annual Conference of the International Speech Communication Association*.

Schuller, B. y Batliner, A. (2013). *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons.

Schuller, B., Steidl, S., y Batliner, A. (2009). The interspeech 2009 emotion challenge. En *Tenth Annual Conference of the International Speech Communication Association*.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., y Narayanan, S. S. (2010). The interspeech 2010 paralinguistic challenge. En *Eleventh Annual Conference of the International Speech Communication Association*.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., y Krajewski, J. (2011). The interspeech 2011 speaker state challenge. En *Twelfth Annual Conference of the International Speech Communication Association*.

Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., Son, R. v., Weninger, F., Eyben, F., Bocklet, T., *et al.* (2012). The interspeech 2012 speaker trait challenge. En *Thirteenth Annual Conference of the International Speech Communication Association*.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., y Narayanan, S. (2013a). Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, **27**(1): 4 – 39. Special issue on Paralinguistics in Naturalistic Speech and Language.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., *et al.* (2013b). The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. En *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

Schuller, B. W., Steidl, S., Batliner, A., Hantke, S., Hönig, F., Orozco-Arroyave, J. R., Nöth, E., Zhang, Y., y Weninger, F. (2015). The interspeech 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition. En *INTERSPEECH*.

Schuller, B. W., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., Elkins, A. C., Zhang, Y., Coutinho, E., y Evanini, K. (2016). The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. En *Interspeech*, Vol. 2016, pp. 2001–2005.

Shobaki, K., Hosom, J.-P., y Cole, R. A. (2000). The ogi kids' speech corpus and recognizers. En *Sixth International Conference on Spoken Language Processing*.

Tomek, I. *et al.* (1976). An experiment with the edited nearest-nieghbor rule.

Vinciarelli, A., Pantic, M., Bourlard, H., y Pentland, A. (2008). Social signals, their function, and automatic analysis: A survey. En *Proceedings of the 10th International Conference on Multimodal Interfaces*, ICMI '08, pp. 61–68, New York, NY, USA. ACM.

Xu, Y. y Goodacre, R. (2018). On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, **2**(3): 249–262.

Yildirim, S., Narayanan, S., y Potamianos, A. (2011). Detecting emotional state of a child in a conversational computer game. *Computer Speech & Language*, **25**(1): 29 – 44. Affective Speech in Real-Life Interactions.

Zhang, Y., Michi, A., Wagner, J., André, E., Schuller, B., y Weninger, F. (2019). A generic human–machine annotation framework based on dynamic cooperative learning. *IEEE Transactions on Cybernetics*, **50**(3): 1230–1239.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., y Schölkopf, B. (2004). Learning with local and global consistency. En *Advances in neural information processing systems*, pp. 321–328.

Zhu, X. y Ghahramani, Z. (2002). Learning from labeled and unlabeled data with label propagation.(2002).

Zhu, X. y Goldberg, A. B. (2009a). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, **3**(1): 1–130.

Zhu, X. y Goldberg, A. B. (2009b). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, **3**(1): 1–130.