UNIVERSIDAD NACIONAL DE SAN AGUSTÍN DE AREQUIPA
ESCUELA DE POSGRADO
UNIDAD DE POSGRADO DE LA FACULTAD DE INGENIERIA DE PRODUCCION Y SERVICIOS

TESIS

UN CORPUS ORAL EN IDIOMA ESPAÑOL CON ACENTO PERUANO PARA EL RECONOCIMIENTO DE EMOCIONES

Tesis presentada por la Bachiller:
ALESSANDRA DANIELA DELGADO MATTOS
PARA OPTAR EL GRADO ACADÉMICO DE
Maestra en Ciencias:Informática,
con mención en Tecnologías de la Información y Comunicación en Gestión y Educación
Asesor: Dr. ALVARO ERNESTO CUNO PARARI
Co-asesor: Mg. WILBER ROBERTO RAMOS LOVÓN

Arequipa, Perú
2022

Quisiera agradecer a nuestro Padre del cielo, a Santa María y San José, quienes me guiaron hacia este gran proyecto llamado Kusisqa, gracias a su sabíduría, a su guía y fortaleza me han permitido finalizar mi labor en este proyecto con este trabajo que espero sirva como modelo para poder generar más corpus del habla en español con acento peruano.

Quisiera agradecer en segundo lugar a mi familia por el constante apoyo durante todo este tiempo alentandome a seguir adelante y cumplir mis metas.

Finalmente agradecer a mis asesores, Dr. Alvaro Cuno y Mag. Wilber Ramos, por apoyarme constantemente en este proyecto que surgió de una idea y que gracias a su constante trabajo y apoyo se convirtió en realidad.

Índice general

Índice de figuras

Índice de cuadros

Agradecimientos

Esta investigación ha sido financiada por el Proyecto Concytec - Banco Mundial "Mejoramiento y Ampliación de los Servicios del Sistema Nacional de Ciencia Tecnología e Innovación Tecnológica" 8682-PE, a través de su unidad ejecutora ProCiencia [Contrato número 014-2019-FONDECYT-BM-INC.INV].

Resumen

[Antecedentes] Los estudios han demostrado que uno de los elementos principales de la comunicación, son las emociones, es por ello, que el rol de ellas ha sido ampliamente analizado en los sistemas automáticos de reconocimiento a partir del habla, los cuáles, requieren además de algoritmos robustos, corpus orales de emociones de calidad.

[Motivación] Los corpus etiquetados emocionalmente son un elemento clave en la implementación de sistemas automáticos de reconocimiento de emociones a partir del habla. La carencia de los mismos hace que los hablantes de un determinado idioma, acento o dialecto, no puedan usufructuar, en toda su amplitud, de los beneficios de este desarrollo tecnológico.

[Objetivos] Esta tesis tiene como objetivo diseñar y construir un corpus de emociones en español con acento peruano, de modo que pueda ser utilizado en el entrenamiento y validación de sistemas basados en *Deep Learning* para el reconocimiento automático de emociones.

[Método] Se comenzó realizando una revisión de la literatura sobre los modelos de clasificación de emociones y los esquemas relacionados al diseño y construcción de corpus orales de emociones. Después, se definieron los criterios para diseñar un método de construcción a partir de audio-vídeos existentes en la plataforma YouTube™. Finalmente, la calidad del corpus fue evaluada de manera cualitativa y cuantitativa.

[Resultados] Se construyó un corpus etiquetado con tres atributos emocionales (valencia, excitación y dominancia), con un tamaño de 7 horas 45 minutos y 52 segundos, contiene voces de un total de 80 participantes (hombres y mujeres en edad adulta) desenvolviéndose en escenarios naturales tales como debates, entrevistas y reportajes.

Asimismo, se encuentra disponible de forma abierta al público en el siguiente enlace: https://zenodo.org/record/5793223#.YczDf2jMLIV

[Conclusión] El corpus oral de emociones fue evaluado cualitativamente en cuanto al alcance, naturalidad, contexto y descriptores, y cuantitativamente en cuanto a su eficacia real, cumpliendo las expectativas para lo que fue construido ya que permitió el entrenamiento y validación en sistemas de reconocimiento de emociones en idioma español con acento peruano, obteniendo un buen desempeño, del 0.84 y 0.73 para las etapas de entrenamiento y validación respectivamente.

Abstract

Background: Emotional corpus are a key element in the implementation of automatic speech emotion recognition systems. The lack of them means that the speakers of a certain language, accent or dialect cannot fully enjoy the benefits of this technological development.

Objective: This thesis aims to design and build an emotional corpus in Spanish with a Peruvian accent so that it can be used in training and validating Deep Learning-based systems for automatic emotion recognition.

Methods: It began by reviewing the literature on emotion classification models and schemes related to the design and construction of speech emotional corpus. Then, the criteria to design a construction method from existing audio-videos on the YouTube™platform were defined. Next, the corpus was built using the proposed design. Finally, the quality of the corpus was evaluated qualitatively and quantitatively.

Results: A corpus labeled with three emotional attributes (valence, arousal and dominance) was built, with a length of 7 hours 45 minutes and 52 seconds, and contains voices of a total of 80 participants (adult men and women) unfolding in natural settings. such as debates, interviews and reports. It is also available without restriction at the following link: https://zenodo.org/record/5793223#.YczDf2jMLIV.

Conclusion: A speech emotional corpus was successfully built, containing qualitative and quantitative features that allow to be used in training and validating *Deep Learning* systems for Spanish with a Peruvian accent emotion recognition. This corpus complements the existing corpus and contributes to raising awareness and promoting the need to create emotional corpus with Latin American accents.

Abreviaturas

ML Machine Learning

DDS Definición estándar del conjunto de datos

DRS Especificación de requisitos del conjunto de datos

DVP Plan de verificación del conjunto de datos

DRR Data Readiness Report

CC Creative Commons

SAM Self Assesment Manikin

IEMOCAP Interactive emotional dyadic motion capture database

EMOTV1 Annotation of real-life emotions for the specifications of multimodal affective interfaces

HEU Emotion A Large-scale Database for Multimodal Emotion Recognition in the Wild

LSSED A Large Scale Dataset and Benchmark for Speech Emotion Recognition

MFCC Coeficientes Cepstrales de Frecuencia Mel

Capítulo 1

Introducción

1.1. Contexto

Las emociones tienen un papel muy importante en la sociedad ya que son un elemento esencial de la comunicación entre las personas. Mediante el habla las personas expresan no solo información verbal sino también emocional. En la actualidad, la comunicación verbal es objeto latente de investigación [26] ya que la emoción en si misma es un concepto muy abstracto y en la comunicación verbal se observan características de entonación, por ejemplo las voces tristes suelen ser de tono bajo y lento mientras que las voces alegres suelen ser todo lo contrario. Así el reconocimiento de emociones en el habla es un contribuyente vital para la interacción humano-computador [53, 91].

Los sistemas de reconocimiento de emociones requieren dos elementos principales [58, 18, 26, 48], los algoritmos de aprendizaje y los corpus. Los corpus son conjuntos de datos, los cuales están compuestos por datos de voz, lenguaje y etiquetas emocionales, los cuales permiten realizar el entrenamiento, validación y pruebas de los algoritmos. Dentro de los algoritmos de aprendizaje, los que vienen recibiendo mayor atención de parte de los investigadores, debido a su disponibilidad, son aquellos basados en el reconocimiento de las emociones a partir del habla [62] utilizando modelos basados en *Deep Learning* [40].

En este contexto, una de las mayores limitaciones para llevar a cabo el estudio de la expresión de las emociones, es la falta de corpus con interacciones genuinas[48],

expresadas por hablantes en entornos reales. Se utilizan corpus actuados u obtenidos, los cuales presentan las emociones de forma controlada y al utilizarlos en aplicaciones de la vida real se dan ciertas limitaciones en el reconocimiento.

Existen tres tipos de corpus orales de emociones, los actuados, los obtenidos y los naturales, que se utilizan para estudiar las emociones basadas en el habla[61]. Uno de los desafíos en el uso de los corpus orales de emociones, es la disponibilidad de corpus orales de emociones naturales[48, 54], ya que son más difíciles de obtener y existe la posibilidad de que algunas emociones estén ausentes y presenten ruido de fondo. Otro desafío es la falta de corpus que contengan voces de hablantes de diferentes grupos etarios [64] y de diferentes idiomas, acentos y dialectos. Ocasionando que la comunidad de investigación no pueda explotar todos los escenarios para poder crear sistemas de clasificación más robustos [15].

En cuanto al factor del idioma, podemos observar que los corpus existentes se pueden agrupar en dos grupos principales: un grupo mayoritario donde encontramos varios corpus de un mismo lenguaje (Inglés y Alemán) [22, 54, 62] y un grupo minoritario compuesto por corpus en lenguas como el Español[21, 64], Hebreo[6], Ruso[51, 50], Koreano [41], Japonés [37, 79], Shindi [43] y Urdu [94].

Es importante recordar que un idioma posee una variedad de dialectos [35], así como los diferentes acentos presentes en los hablantes [92]. Por ejemplo, no es lo mismo comparar a un hispano hablante español que un hispano hablante peruano. Los estudios realizados por [35] en cuanto al dialecto y [92] en cuanto al acento, han demostrado que la variación en el habla debido al dialecto es un factor que afecta significativamente al sistema del habla ya que cada dialecto se diferencia por características acústicas como la fonética de las vocales y consonantes, el ritmo y las características prosódicas[1], también la gramática y el vocabulario. Siendo estás características las que suelen generar un desajuste en el rendimientos de los sistemas automáticos de reconocimiento de voz.

Para mejorar el rendimiento de los sistemas de reconocimiento de emociones, en esta tesis, se plantea el diseño y construcción de un corpus oral en español con acento peruano para que los sistemas de reconocimiento de emociones, basados en modelos de

[1]Como el acento, la entonación, el tono, el ritmo, la melodía, las pausas, la velocidad de elocución y la calidad de la voz.

Deep Learning, puedan procesar las características lingüísticas y prosódicas de nuestra lengua y hablantes de modo que se pueda entrenar y validar el sistema y así mejorar la confiabilidad de los mismos.

El corpus propuesto es de libre acceso, ha sido obtenido a partir de escenarios realistas de interacción entre hispano hablantes con acento peruano. Contiene segmentos de audio etiquetados con tres atributos emocionales: valencia (de negativo a positivo), excitación (de calma a activo) y dominancia (de dominado a dominante).

Estas anotaciones dimensionales permiten el estudio de las emociones más allá de las básicas ya que mediante esta escala se puede determinar el estado afectivo[2] del participante [25].

El corpus ha sido evaluado cualitativamente y cuantitativamente. En cuánto a la evaluación cualitativa se obtuvieron los siguientes resultados: la diversidad emocional del corpus, basada en la cantidad de segmentos por cada dimensión emocional, donde la dominancia fue la que obtuvo los mejores resultados, ya que predomina la dimensión "algo dominante" sobre la "neutra". Además fue utilizado el coeficiente *Alpha Cronbach* para medir la concordancia entre los valores asignados por los tres etiquetadores del corpus, el cuál obtuvo un valor mayor de 0.9. En la discusión, se demostró la satisfacción de los cuatro factores principales que deben ser considerados en el diseño de los corpus, según la definición realizada en [22].

En cuanto a la evaluación cuantitativa se obtuvieron los siguientes resultados en base a la evaluación realizada del corpus en dos arquitecturas de deep learning, una red neuronal convolucional y una red neuronal residual, obteniendo los resultados para la etapa de entrenamiento de 0.842 y 0.775 respectivamente (promedio de los resultados de las tres dimensiones emocionales) y para la etapa de validación de 0.726 y 0.654 respectivamente.

Los sistemas de reconocimiento de emociones pueden ser aplicados en sistemas críticos como en un sistema de conducción de automóviles a bordo [97], utilizando la información del estado mental del conductor para mantenerlo alerta durante la con-

[2]Las emociones y actitudes forman parte del estado afectivo, las actitudes son creencias, preferencias y predisposiciones relativamente duraderas y coloreadas afectivamente hacia objetos o personas, estas son menos intensas y tienen una mayor duración que las emociones. Los estados afectivos se caracterizan por mostrar el grado de confianza o entusiasmo del hablante más no se caracterizan por una valencia positiva o negativa, como si la tienen las emociones.

ducción, ayudando a evitar algunos accidentes causados por el estrés del conductor. Los médicos también pueden utilizar los contenidos emocionales del habla de sus pacientes como herramienta de diagnóstico de diversos trastornos [82, 19]. Las conversaciones de las centrales de llamadas se pueden utilizar para analizar el estudio del comportamiento de los asistentes de llamadas con sus clientes ayudando a mejorar la calidad del servicio, así como el análisis de las llamadas en servicios de emergencia ayudaría a evaluar la autenticidad de las solicitudes.

También pueden ser aplicados para sistemas de aprendizaje y entretenimiento, como las aplicaciones de películas interactivas, narración de historias y E-tutoría, las cuales serían más prácticas si podrían adaptarse al estado emocional de los oyentes o estudiantes.

El corpus diseñado y construido en esta tesis es un corpus natural, ya que los participantes de cada uno de los audios, cuidadosamente seleccionados, de diálogos, entrevistas y conversaciones, se expresan libremente. Además de ser un corpus natural, es un corpus en idioma español y con acento peruano, lo que permite representar las características acústicas y prosódicas, gramática y vocabulario, propias del español con nuestro acento. Debido a que un idioma posee una variedad de dialectos, en el presente corpus, se encuentra la presencia de los dialectos estándar y vernacular. Uno de los factores, que afectan en el rendimiento de los sistemas automáticos de reconocimiento de emociones a partir del habla, es el género [42, 28], ya que la expresión de emociones tanto en hombres como mujeres, no es igual, por lo que, este corpus, presenta la participación, tanto del género femenino como masculino, de forma equilibrada. Debido a que se ha diseñado y construido un corpus oral de emociones natural, se optó por utilizar las dimensiones emocionales de valencia, excitación y dominancia, las cuales, representan de forma más adecuada el contenido emocional del corpus. Finalmente el corpus se evaluó en dos modelos de deep learning, con el propósito de medir la capacidad del corpus y determinar si permite el reconocimiento de emociones. Demostrando que el corpus tanto para las etapas de entrenamiento y validación, obtiene un resultado considerablemente bueno. Contribuyendo así, a que otros modelos de deep learning puedan utilizarlo, identificando la expresión de emociones en peruano-hablantes y así incentivar la investigación del manejo de las emociones en los sectores de educación y

salud en nuestra sociedad.

1.2. Motivación y/o Justificación

Es fundamental señalar que la construcción de un corpus especializado para la detección de emociones es una tarea compleja [93] ya que requiere del esfuerzo de equipos de actores, transcriptores, anotadores y evaluadores expertos y previamente capacitados, lo que conlleva un coste, no sólo económico sino también muy elevado en el tiempo. Aunque esta modalidad genera expresiones exageradas que difieren sutilmente de los comportamientos observados durante las interacciones en la vida diaria [49]. Frente a esto, una alternativa es crear un corpus natural con contenido emocional equilibrado.

Asimismo, la variación en el habla debido al dialecto es un factor que afecta significativamente al sistema del habla ya que cada dialecto se diferencia por características acústicas como la fonética de las vocales y consonantes, el ritmo y las características prosódicas, así como también la gramática y el vocabulario. Es por ello que el corpus, abarca dos tipos de dialectos del idioma español con acento peruano, el estándar y vernacular, con una cantidad de participantes equilibrado entre hombres y mujeres, factor que también influye en el reconocimiento de las emociones.

Por las razones anteriormente expuestas se justifica un corpus oral en español pero con acento peruano. Diferentes sistemas de reconocimiento automático de emociones, como los basados en *Hidden Markov Model*, y algoritmos de *Deep Learning*, que han logrado resultados notables [78], podrían utilizar el corpus para obtener una mejora en la precisión y en su robustez al momento de reconocer las emociones del habla de los hablantes peruanos. Tal como se ha demostrado validando el corpus en dos modelos de deep learning, obteniendo buenos resultados en el reconocimiento de las emociones. Del mismo modo también podría utilizarse en las investigaciones de reconocimiento de emociones del habla multilingües [44, 8, 96].

1.3. Definición del problema

La carencia de un corpus oral de emociones en español con acento peruano implica que no se puedan construir sistemas robustos frente a diferentes escenarios lingüísticos y que los hablantes del mismo no puedan usufructuar, en toda su amplitud, de los beneficios del avance tecnológico de los sistemas de reconocimiento automático de emociones a partir del habla.

1.4. Objetivos

1.4.1. Objetivo general

Diseñar y construir un corpus oral en idioma español con acento peruano, etiquetado emocionalmente, que permita realizar el entrenamiento y validación en sistemas de reconocimiento de emociones basados en *Deep Learning*.

1.4.2. Objetivos específicos

- Definir los criterios para diseñar un método de construcción a partir de audiovídeos existentes en la plataforma YouTube™.
- Construcción del corpus oral en idioma español con acento peruano.
- Realizar una evaluación cualitativa y cuantitativa en cuánto a la eficacia del corpus construido.
- Validar el corpus, en modelos de deep learning, de modo que se pueda demostrar que permite el reconocimiento de emociones.

Capítulo 2

Marco teórico

2.1. Ingeniería de corpus

En la década de 1950, se introdujo el término paralingüística, que significa, junto a la lingüística, el cuál, explica el cómo se dice algo. La paralingüística, se ocupa del habla y el lenguaje, que son principalmente, medios de comunicación.

El estudio del habla y la emoción, se remonta a las primeras décadas del siglo XX [29], en las cuáles se realizaron análisis sobre las expresiones del habla, relacionadas con las diversas emociones, como el miedo, enojo, desprecio, pena, indiferencia, aburrimiento, interés, placentero, no placentero, autoritario, sumiso, entre muchas otras, lo que permitió identificar características distintivas por cada emoción en el habla, permitiendo el estudio entre las dimensiones de la emoción y los parámetros del habla.

A mediados de la década de 1990, se dieron los primeros pasos hacia el procesamiento automático de emociones en el habla, dando paso a la creación de los corpus orales de emociones, los cuáles requieren datos de voz y lenguaje, junto con información de etiquetas, todo esto forma parte del estudio de la paralingüística computacional [74].

El procesamiento automático del habla, abordó los fenómenos más allá del reconocimiento puramente de la voz. Sin embargo la mayoría de enfoques automáticos de reconocimiento de emociones en el habla, hasta hace unos años, estaban representados en su mayoría por datos actuados, grupos homogéneos de hablantes, algunas caracte-

rísticas prosódicas [1] y clases de emociones muy pronunciadas, como alegría, tristeza, miedo, enojo, desagrado y neutral [23]. Este enfoque, ha cambiando en los últimos años, ya que el objetivo, es representar la realidad máxima [55], por lo que se deben usar todos los datos disponibles, lo que quiere decir, el tener que lidiar con elementos ambiguos.

La ingeniería del corpus, involucra todos los pasos a seguir antes de que los datos de voz puedan ser procesados [13]:

1. Decidir, si corresponde utilizar grabaciones existentes, como, TV, transmisión, internet o el diseño de un escenario de grabación, que podría ser un laboratorio o adecuar una oficina, la cuál podría o no contar con agentes virtuales o robots.

2. Seleccionar el tipo de discurso, que podría ser, leído, incitado, actuado, suscitado o realista. Así como decidir el tipo de grabación (micrófono de conversación cercana o de sala, grabaciones de video, en un entorno multimodal).

3. Reclutar el personal necesario, como supervisores, en el caso de seleccionar un escenario de Mago de Oz, considerar las regulaciones de privacidad apropiadas para las grabaciones y si es necesario, la transferencia a medios de almacenamiento.

4. Transliteración, es decir, la transcripción ortográfica de los datos, que puede incluir la anotación de eventos extralingüísticos o no lingüísticos.

5. Definición y extracción de unidades de análisis apropiadas como palabras, fragmentos, giros, movimientos de diálogo con criterios apropiados o basados en criterios físicos, como pausas del habla, partición de archivos de voz en n segmentos de igual longitud.

6. Establecer fenómenos para ser anotados y procesados, seleccionar la anotación de estados o rasgos sobre una base categórica o continua, poder establecer un patrón de oro para las anotaciones, basado, si es posible, en varios anotadores, y evaluar la calidad de estas anotaciones, aplicando algunas medidas de correlación/correspondencia.

7. El uso de otros pasos de preprocesamiento, como procesamiento manual o la corrección de valores de características procesados automáticamente.

8. Contar con la documentación detallada de las condiciones de grabación acústica

[1] La prosodia es un conjunto de rasgos suprasegmentales en los que se incluyen el acento, la entonación, el tono, el ritmo, la melodía, las pausas, la velocidad de elocución y la cualidad de la voz [47]

de la sala y todos los demás detalles que podrían ser relevantes. Realizar pruebas de percepción si es correspondiente.

9. Definir la división de datos en conjuntos de entrenamiento, desarrollo y prueba y finalmente, de ser posible, liberar gratuitamente los datos, teniendo en cuenta las condiciones de privacidad.

Para el reconocimiento automático de voz [85], las transcripciones ortográficas eran suficientes, pero, actualmente, se necesitan más datos para poder tener un buen rendimiento de dichos sistemas, es por ello, que, para el habla se puede aprovechar el contexto lingüístico como extralingüístico[2]. Además, la anotación, se realiza asignando etiquetas a las unidades de voz, este proceso se puede realizar, basado en modelos, datos o aplicaciones, en su mayoría se utiliza la combinación de estos tres enfoques.

Los modelos, de tipo *Gran N*, prevalecen para la personalidad y la emoción, este modelo, es utilizado, para los rasgos de personalidad, porque se basa en un modelo genérico de características de larga duración, sin embargo, es una mala elección para las emociones, si se trata con datos realistas, no actuados, debido a que los participantes, son libres de elegir su propio estado emocional, por lo tanto, las anotaciones, en estos casos debieran basarse en datos, seleccionando aquellos estados o rasgos para la anotación, que sean de interés para el estudio realizado. Por ejemplo, en un escenario, de centro de llamadas, solo interesaría saber si el cliente está enojado y cuando lo está. Por su puesto, mientras mayor información adicional se tenga, será más probable de emplear dicha información, para modelar el fenómeno que realmente es de interés para el estudio, saber el sexo, idioma, edad, estado mental o emocional anterior, puede aprovecharse, para modelar y procesar otros estados o rasgos.

Las dimensiones, se pueden anotar con escalas de calificación o de forma continua, las anotaciones, pueden ser, atribuciones categóricas binarias (Sí/No) o características n-arias o atribuciones de opción múltiple. El concepto de escalas de calificación, se remonta a la escala de actitud propuesta en [46]. Los pros y contras del modelado de anotaciones categóricas versus dimensionales/continuas se han discutido ampliamente, desde el punto de vista de la investigación básica, se favorecen a las anotaciones más

[2]Que es exterior a la lengua, aunque influye en el proceso global de la comunicación. Por ejemplo: "la gesticulación es un factor extralingüístico de gran importancia para la comunicación".

detalladas y desde el punto de vista de la investigación orientada a la aplicación, se prefieren solo las etiquetas para aquellos fenómenos que más interesan en el estudio realizado.

A menudo en paralingüística computacional, el estándar de oro, no es confiable, es decir, la etiqueta de entrenamiento y prueba en sí misma puede ser errónea, ya que depende en gran medida de la tarea, por ejemplo, generalmente se conoce la edad de un participante, pero la emoción del mismo puede ser difícil de evaluar.

Al interpretar los resultados, se debe tener en cuenta, que la referencia suele ser el patrón de oro y no necesariamente la verdad fundamental (verdad real medida sobre el terreno). Esto tiene un doble impacto: por un lado, los modelos aprendidos son propensos a errores, por otro lado, los resultados de la prueba pueden ser interpretaciones excesivas o insuficientes. Por lo tanto para lograr un estándar de oro confiable, por lo general, se utilizan varios anotadores, cuanto menos segura es la tarea, mayor es el número de anotadores.

Hay un par de medidas para identificar el acuerdo entre etiquetadores (confiabilidad entre evaluadores), en el caso habitual que participan dos o más anotadores. Si la tarea se modela continuamente, como la simpatía de un hablante, en una escala continua, el coeficiente de correlación (media) y el error lineal medio (MLE) (promedio), el error absoluto medio (MAE) o el error cuadrático medio (MSE) y la desviación estándar, entre los etiquetadores, se utilizan con frecuencia.

Por otro lado, se puede preferir el MSE o similar, si la tarea, tiene un punto de referencia bien definido y una base sólida, como en el caso de la determinación de la altura o la edad del hablante, ya que esto, puede ser más intuitivo al momento de interpretarlo.

En el caso del modelado categórico, se puede emplear una variedad de medidas para la evaluación del acuerdo, como el *alfa de Krippendorff*, o el *kappa de Cohen* o *Fleiss*. Un modelo continuo, también se puede discretizar, por lo que, las últimas estadísticas, también se pueden usar en este caso, a menudo con una ponderación lineal o cuadrática. El coeficiente de correlación de *Pearson*[63] y el coeficiente de correlación de rangos de *Spearman rho* [80], se pueden usar para tales intervalos clasificados.

En [34], afirman, que, las propiedades acústicas del habla emocional, se capturan

mejor, utilizando modelos formados a partir de evaluaciones promediadas, en lugar de evaluaciones individuales específicas.

Actualmente, existen nuevos métodos de anotación, que implican un costo reducido, como son los métodos de comunidad o anotación distribuida, como crowdsourcing, por ejemplo, por *Amazon Mechanical Turk*[3] o los métodos de *Models in the loop*, como los proyectos *Dynabench Data Collection and Benchmarking Platform & Mephisto Dataset Collection Tool*, los cuáles, soportan la realización de múltiples tareas en bucles, también permiten el esfuerzo en comunidad con propietarios expertos en las tareas, que toman las decisiones y la recopilación de datos de forma dinámica, de modo, que los sistemas automáticos de recomendación, no se queden en el caso promedio, sino, que puedan mejorar continuamente, validándolos con el peor de los casos. Además, se han realizado estudios sobre la fusión beneficiosa de bases de datos [75], que permiten obtener aún mayores cantidades de datos, sin la cantidad habitual de esfuerzo de anotación.

También, se demostró, que el material de entrenamiento sintetizado[76] es muy beneficioso en las pruebas de corpus cruzados, es decir, utilizar una base de datos diferente para el entrenamiento y la validación.

Según, el análisis de nueve bases de datos realizado en [13], se puede observar, que la mayoría de los datos son solicitados en lugar de espontáneos. En particular, para tareas paralingüísticas menos investigadas, prevalecen las bases de datos con lenguas germánicas y latinas. Estos dos problemas, del dominio del material solicitado y por lo tanto fonéticamente limitado y el desequilibrio de los idiomas presentados, resaltan la gran necesidad de una mayor diversidad en el aspecto lingüístico.

Del mismo modo, las condiciones de laboratorio prevalecen durante la grabación, además de los datos tomados de los medios de difusión. El tamaño de las bases de datos es sorprendentemente pequeño, a menudo, los conjuntos contienen unos pocos cientos hasta unos pocos miles de instancias o unas pocas horas de material de voz. Esto contrasta, marcadamente, con campos relacionados como el reconocimiento de voz, donde se utilizan hasta varios años de material de voz para entrenar y probar sistemas.

Finalmente, en un análisis que realizaron sobre el desempeño del reconocimiento

[3]https://www.mturk.com/

automático paralingüístico, se pudo observar, que se obtuvo el mejor resultado para tareas de dos clases, sin embargo, para un mayor número de clases, el número baja, lo que demostró, que mientras más subjetiva es una tarea más desafiante es.

Luego de revisar los componentes involucrados en el diseño y construcción de un corpus, ahondaremos en cada uno de ellos, para poder distinguir su desempeño en el corpus. Las funciones, destrezas y habilidades de la voz, el habla y el lenguaje [57] están relacionadas. Algunos textos usan los términos casi indistintamente, pero para los científicos y profesionales médicos es importante distinguirlos.

2.2. Voz

La voz, es el sonido que producen los humanos utilizando los pulmones y las cuerdas vocales de la laringe. Sin embargo, la voz no siempre se produce como habla, ya que los bebés balbucean y los adultos ríen, cantan y lloran. La voz, se genera por el flujo de aire de los pulmones a medida que las cuerdas vocales se acercan. Cuando el aire pasa por las cuerdas vocales con suficiente presión, las cuerdas vocales vibran. La voz es tan única como la huella digital, ayuda a definir la personalidad, estado de ánimo y salud. La voz consta de tres componentes importantes, el tono, el volumen y la calidad. El tono, es el tono alto o bajo de un sonido basado en la frecuencia de las ondas sonoras. La sonoridad es el volumen percibido (o amplitud) del sonido, mientras que la calidad se refiere al carácter o atributos distintivos de un sonido.

2.3. Habla

Los seres humanos, pueden expresar sus pensamientos, sentimientos e ideas oralmente entre sí, a través de una serie de movimientos complejos que alteran y moldean el tono básico creado por la voz en sonidos específicos y decodificables. El habla, se produce mediante acciones musculares coordinadas con precisión en la cabeza, el cuello, el pecho y el abdomen. El desarrollo del habla es un proceso gradual que requiere años de práctica.

Según [24], la señal del habla, es el método de comunicación más rápido y natural entre las personas. Este hecho ha motivado a los investigadores a pensar en el habla

como un método rápido y eficiente de interacción entre humanos y máquinas. Sin embargo esto requiere que la máquina tenga la inteligencia suficiente para reconocer las voces humanas.

Desde finales de los años cincuenta, ha habido una gran investigación sobre el reconocimiento de voz, que se refiere, al proceso de convertir el habla humana en una secuencia de palabras, sin embargo, a pesar de todos estos avances aún nos encontramos lejos de tener una interacción natural entre el hombre y la máquina, porque la máquina no comprende el estado emocional del hablante. Lo que ha introducido un campo de investigación, el reconocimiento de emociones del habla, que se define como la extracción del estado emocional de un hablante a partir de su habla. El objetivo principal de emplear el reconocimiento de emociones del habla es adaptar la respuesta del sistema al detectar frustración o molestia en la voz del hablante.

Las primeras investigaciones de reconocimiento de emociones a partir del habla, fueron conducidas a mediados de los años ochenta, utilizando propiedades estadísticas de ciertas características acústicas [84].

2.4. Lenguaje

El lenguaje, es la expresión de la comunicación humana a través de la cual se pueden experimentar, explicar y compartir el conocimiento, las creencias y el comportamiento. Este intercambio esta basado en signos, sonidos, gestos o marcas sistemáticas y de uso convencional que transmiten significados comprendidos dentro de un grupo o comunidad.

2.5. Las emociones

Según[54], la naturaleza de los diferentes tipos de emociones así como la falta de dimensiones consensuadas, debido a la dificultad para distinguir los diferentes estados emocionales por sus intensidades y duraciones relativas, como por ejemplo, los estados de ánimo, las actitudes, las posturas interpersonales e incluso rasgos de personalidad afectiva, han hecho que sea complejo definirlas, sin embargo algunos autores han podido clasificarlas.

Referencia	Emociones Básicas
[52]	enojo, júbilo, disgusto, sujeción, miedo, asombro, ternura
[9]	tristeza, enojo, coraje, miedo, aversión, deseo, abatimiento, desesperación, esperanza, odio, amor
[38]	sorpresa, desprecio, disgusto, enojo, culpa, angustia, interés, alegría, miedo, vergüenza
[67]	aceptación, enojo, miedo, tristeza, anticipación, sorpresa, disgusto, alegría
[23]	enojo, alegría, tristeza, disgusto, sorpresa, miedo
[33]	rabia y terror, ansiedad, alegría
[83]	interés, alegría, enojo, desprecio, angustia, disgusto, miedo, sorpresa, vergüenza
[89]	tristeza, alegría
[31]	interés, deseo, sorpresa, alegría, asombro, dolor
[60]	tristeza, ansiedad, alegría, enojo, disgusto

Cuadro 2.1: Resumen de las emociones básicas utilizadas por los investigadores. [54]

[71], Definió las emociones, como episodios de variaciones coordinadas en muchos componentes, incluyendo reacciones neurofisiológicas, expresión motora, sentimiento subjetivo y el proceso cognitivo en respuesta a eventos y estímulos exteriores e interiores.

La investigación relacionada a las emociones está basada en dos supuestos: el primero, es el número discreto de emociones básicas que son representativas para todas las otras emociones y el segundo, considera la continuidad de las emociones e intenta mapear todas las emociones en un espacio dimensional o tridimensional.

El cuadro 2.1, muestra el resumen de los grupos de emociones básicas utilizadas por los investigadores.

2.6. Dimensiones emocionales

Según [54], el segundo supuesto, en el cuál se considera la continuidad de las emociones, la caracterización de las emociones se puede ver como un modelo de espacio continuo de tres clases, al que se le hace referencia como espacio de excitación, valencia y poder. En este modelo las emociones activas y pasivas están representadas por la dimensión de excitación, las emociones positivas y negativas, es decir de la ira a

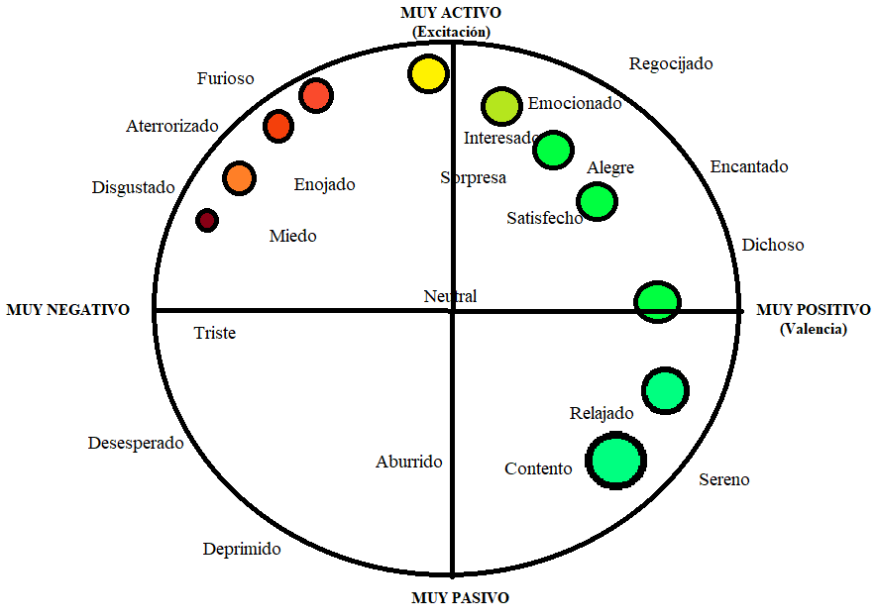


Figura 2.1: Modelo espacial de emociones (dos dimensiones). Fuente [54].

.

la felicidad, están representadas por la dimensión de valencia y la dimensión de poder, representa el sentido o grado de poder de control sobre las emociones de uno. La figura 2.1, muestra un número de estas emociones distribuidas es un espacio de dos dimensiones.

2.7. Corpus orales de emociones

En la literatura, las fuentes de datos a partir del habla son llamadas, conjuntos de datos [22] o corpus, en lugar de bases de datos, ya que son colecciones de datos en menor escala, generalmente creados para examinar una cuestión particular que no se encuentra disponible abiertamente al público.

Para la construcción de un corpus, es necesario contemplar las siguientes cuestiones: el alcance, la naturalidad, el contexto del contenido y los tipos de descriptores apropiados, los cuáles se explicarán a continuación.

2.7.1. El alcance

El término alcance, se utiliza para cubrir los diferentes tipos de variaciones que un corpus o conjunto de datos podría incorporar, cómo, el número de distintos hablantes, el lenguaje hablado, el género de los hablantes, tipos de estados emocionales considerados, tokens de un estado dado, el entorno social. Las comparaciones entre lenguas y culturas son limitadas, pero sugieren diferencias considerables, por ejemplo en la sociedad Japonesa una demostración abierta de las emociones puede considerarse un comportamiento antisocial o egoísta, sin embargo se considera normal sonreír cuando se está enojado o avergonzado. El género también es una variable socio-lingüística clave. Además, según los análisis realizados en la literatura se ha hallado que un escenario normal para la expresión de emociones es el diálogo.

Un segundo aspecto del alcance, se relaciona con el rango de emociones consideradas, según las investigaciones realizadas en el artículo [22], mencionan que el alcance emocional de los corpus debe pensarse detenidamente, ya que es probable que el alcance tuviera que ser muy amplio, por lo que se debe establecer un conjunto de puntos de referencia que sea apropiado para la investigación del habla, lo cual es una tarea empírica, que en si misma depende del acceso a datos que abarcan la gama conocida de estados emocionales.

2.7.2. La naturalidad

El precio de la naturalidad es la falta de control, la emoción tiene una imprevisibilidad que dificulta la recolección de muestras de personas en un estado objetivo, ya sea inducida o espontánea. Por lo que, el objetivo de la investigación juega un papel muy importante, ya que en algunos casos, la naturalidad puede no ser el objetivo relevante.

2.7.3. El contexto del contenido

Existe una evidencia directa entre el contexto y los oyentes, ya que lo utilizan para determinar la significancia emocional de las características vocales. Se pueden distinguir cuatro tipos de contexto:

- Semántico: Existe un claro potencial para la interacción entre el contenido y las

señales vocales.

- Estructural: Parece probable que muchos signos de emoción se definan en relación con las estructuras sintácticas: patrones de estrés, patrones de entonación predeterminados, etc. Se observa con menos frecuencia la posibilidad de que la emoción pueda ser señalada por variaciones en el estilo, que se expresan en características estructurales de los enunciados (frases largas o cortas, repeticiones e interrupciones, etc.).

- Intermodal: El hecho, de que se pueda comunicar una amplía gama de emociones vía telefónica, demuestra que el análisis que se ocupa exclusivamente del habla es razonable. Sin embargo el habla funciona a menudo como un suplemento de información sobre las emociones en lugar de una fuente independiente, ya que normalmente vemos y oímos al hablante.

- Temporal: El habla natural implica patrones distintivos de cambio a medida que la emoción fluye a lo largo del tiempo.

2.7.4. Los descriptores

Construir un corpus requiere, por un lado, técnicas para describir el contenido lingüístico y emocional y por otro lado el habla. El etiquetado preciso de las emociones en un corpus actuado puede ser más fácilmente realizado mediante emociones categóricas, pero el de un corpus oral natural, es más complejo, ya que involucra la gradación hacia adentro y hacia afuera de los picos emocionales.

2.7.5. Accesibilidad

Adicionalmente, se considera el siguiente factor relevante al momento de diseñar y construir un corpus ya que el valor de una corpus incrementa enormemente si esta disponible a toda la comunidad, así no hay necesidad de duplicar el esfuerzo, los algoritmos pueden ser comparados con los mismos datos y más.

Dos cuestiones principales influyen la disponibilidad: el formato y la ética. El formato de los datos necesita ser estándar o transparente, no solo para la materia prima

sino también para los descriptores utilizados. El otro problema fundamental, es el de la ética y los derechos de autor, particularmente, con los datos obtenidos naturalmente o de fuentes naturales como la radio y la televisión, ya que generalmente se habla de temas personales de los participantes y porque presenta restricciones en cuanto a los derechos de autor.

2.8. Machine learning

Según [59], el componente principal de los sistemas de ML son los datos. ML es un subconjunto de la inteligencia artificial. El ML, está definido como los sistemas de computadoras utilizados para desarrollar la tarea sin una instrucción específica pero basada en datos de ejemplo y en experiencias pasadas. A más entrenamiento con datos de buena calidad, estos sistemas pueden resolver tareas basados en los datos entrenados en vez de utilizar la intuición y estos sistemas pueden adoptar nuevas situaciones basados en nuevos datos entrenados.

2.8.1. Algoritmos de machine learning

Los algoritmos de ML, son programas que son entrenados y analizados con datos de entrada para predecir una salida con un rango adecuado, mientras que estos algoritmos son entrenados con nuevos datos, este aprende y mejora sus operaciones para alcanzar el desempeño y desarrollar inteligencia sobre el tiempo.

En la figura 2.2 podemos apreciar los tipos de algoritmos de ML utilizados comúnmente:

- Aprendizaje supervisado: En este tipo de aprendizaje, el algoritmo crea una función que se puede conectar con las entradas de las salidas deseadas.
- Aprendizaje no supervisado: En este tipo de aprendizaje, el objetivo principal para el sistema es aprender como hacer algo nuevo sin entrenar el sistema con datos nuevos.
- Aprendizaje semi supervisado: Este tipo de aprendizaje incluye ambos tipos de

datos, etiquetados y no etiquetados para generar una función o clasificador apropiado.

- Aprendizaje de refuerzo: Este es un tipo de aprendizaje donde los sistemas aprenden y se adaptan a una situación dada observando el mundo en tiempo real.
- Transducción: Este aprendizaje se asemeja más al aprendizaje supervisado pero el sistema intenta predecir una salida diferente basado en las entradas, los resultados del entrenamiento y las nuevas entradas.
- Aprendiendo a aprender: En este aprendizaje el algoritmo aprende sesgos inductivos de sus experiencias previas.

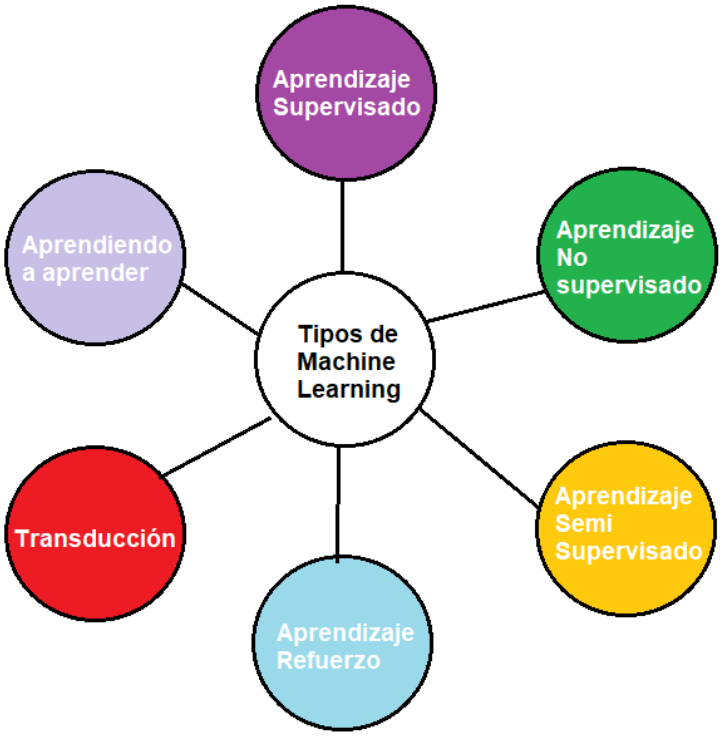


Figura 2.2: Tipos de algoritmos de ML. Fuente [59].

.

2.8.2. Aprendizaje profundo

Las redes neuronales son algoritmos de ML, de aprendizaje supervisado, los cuáles están inspirados en las conexiones de las neuronas en el cerebro. Una Red Neuronal,

tiene como objetivo aproximar una función $y = f(x)$, es decir que debe mapear una entrada x a una categoría y. La red neuronal aproxima esta función de la forma $y = f(x, \theta)$ donde θ son los parámetros que tienen que ser aprendidos para mejorar la función de aproximación.

Las redes neuronales suelen estar formadas por varias funciones, de la forma: $f(x) = f^3(f^2(f^1(x)))$. Las funciones son llamadas capas. Siendo f^1 la primera capa y así sucesivamente. La cantidad de capas se conoce como profundidad, es por ello, que deriva el nombre de aprendizaje profundo. La red neuronal, debe buscar minimizar una función de coste o error, para lo cuál, el proceso de entrenamiento esta ligado al conjunto de datos de entrenamiento el cuál está conformado por pares (x, y), donde x, es la entrada de la red neuronal y y es la salida esperada.

- Redes neuronales convolucionales: Estas redes, se basan, en la idea de tener filtros que puedan extraer las características más resaltantes de una imagen. Anteriormente, estos filtros eran introducidos manualmente, sin embargo, actualmente pueden ser calculados automáticamente usando el algoritmo *back propagation*. Las redes convolucionales, están conformadas, por varias capas convolucionales, las cuáles aplican un conjunto de filtros de convolución sobre una entrada y al resultado se le aplica una función de activación, generalmente, la función $RELU$.

- Redes neuronales residuales: Son una especialización, que busca explotar la repetición de patrones en una serie temporal, utilizan la salida de la red en el tiempo t como entrada de la próxima iteración en el tiempo $t + 1$. Estas redes suelen sufrir de un problema conocido como desvanecimiento de la gradiente, en las capas anteriores, cuando se tienen muchas capas. Por este motivo, se propone, al tener una red con x capas, que funcione correctamente, añadir más capas, con el mapeo de la salida de las x capas, que funcionan correctamente.

  La ventaja, es que las funciones residuales, al utilizar pesos que tienden a 0, dichas funciones se convierten en funciones identidad, ya que la salida y, es igual a la entrada x. Por lo que al ir añadiendo bloques, no modificaría el comportamiento de la red.

2.8.3. Métodos de evaluación

La elección de la función de pérdida, es un aspecto crítico para los sistemas de *Deep Learning*, ya que, dicha función, busca minimizar la función de coste, permitiendo un mejor desempeño en la predicción realizada por la red neuronal. Por ello, a continuación, se explicarán las funciones basadas en error y correlación que han presentado un mejor desempeño en la literatura [10].

Error-based Loss Function

Un error cuadrático medio, para medir la desviación entre el grado de predicción de la emoción x y la etiqueta estándar de oro y esta dada por:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$$

donde n, es el número de medición (calculado por el tamaño de lote) para las tres dimensiones (valencia, excitación y dominancia):

$$MSE_T = MSE_V + MSE_A + MSE_D$$

Se añadieron unos factores de ponderación, para la valencia y excitación. Por lo tanto el MSE total se convirtió:

$$MSE_T = \alpha MSE_V + \beta MSE_A + (1 - \alpha - \beta)MSE_D$$

donde α y β son factores de ponderación para la valencia y arousal. El factor de ponderación, para la dominancia se obtiene restando 1 a las dos variables.

Correlation-based Loss Function

CCC, es una métrica común, en el reconocimiento de dimensiones emocionales, para medir el acuerdo entre la verdadera dimensión de la emoción con el grado de emoción predicho. Si las predicciones cambiaron de valor, la puntuación se penaliza en proporción a la desviación. *CCC* está formulada como:

$$CCC = \frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

donde ρ_{xy}, es el coeficiente de correlación de Pearson entre x i y, σ, es la desviación estándar y μ, es el valor medio. El rango de CCC va de -1 (perfecto desacuerdo) a 1 (perfecto acuerdo). Por lo tanto, la función de pérdida *CCC (CCCL)*, para maximizar la concordancia entre el verdadero valor y la emoción predicha se puede definir como:

$$CCCL = 1 - CCC$$

Similar al aprendizaje multitarea en MSE, se acomodan las funciones de pérdida de valencia ($CCCL_V$), excitación ($CCCL_A$) y dominancia ($CCCL_D$). $CCCL_T$ es una combinación de las tres funciones de pérdida:

$$CCCL_T = \alpha CCCL_V + \beta CCCL_A + (1 - \alpha - \beta)CCCL_D$$

donde α y β, son los pesos de ponderación para cada función de pérdida de la dimensión emocional. Los mismos factores de ponderación son utilizados para *MSE* y CCC ($\alpha = 0{,}1$ y $\beta = 0{,}5$).

2.9. Importancia de la calidad de los datos para sistemas de ML

Según [3], se introdujo el concepto de Data Readiness Report (DRR) como documentación que acompaña al corpus, la cual, permite a los consumidores de los datos obtener percepciones detalladas en la calidad de los datos de entrada. DRR, sirve como registro de todas las operaciones de evaluación de los datos así como las transformaciones aplicadas. Captura y documenta las acciones tomadas por varias personas en un flujo de trabajo de preparación y evaluación de los datos.

Los datos que ingresan al proceso de un sistema de ML están sujetos a un procesamiento previo por varias partes interesadas, por lo que, al utilizar esta herramienta permite la reutilización incrementando los índices de productividad de los profesionales

de datos, debido a que ellos gastan un buen porcentaje de su tiempo en la preparación de los datos. El informe de preparación de datos, es como una evaluación adjunta, que permite a los usuarios de los datos obtener conocimiento sobre la calidad de los datos de entrada en varias dimensiones, tener una documentación completa de todos los datos, propiedades y problemas de calidad, incluidas las operaciones de datos de varias personas para obtener un registro detallado de cómo han evolucionado los datos.

En el siguiente estudio [59], también se valora la importancia de los datos para los sistemas de ML,ya que sostiene que la calidad de los datos de salida en ML, depende de los datos de entrada, siendo los datos de entrada de buena calidad cruciales para el resultado de los sistemas de ML.

En este trabajo, proponen un modelo de calidad de datos, basado en las experiencias industriales de un científico de datos, para poder identificar datos de calidad para el aprendizaje automático.

El trabajo se divide en dos aspectos:

- Se realizó una revisión en la literatura de diferentes corpus, para identificar los atributos y modelos de calidad de datos para el aprendizaje automático.
- Una vez obtenidos los atributos y modelos de calidad, en el estudio anteriormente mencionado, se realizaron entrevistas a quince científicos de datos de múltiples locaciones, con los atributos seleccionados en el estudio y en base a los resultados obtenidos en las entrevistas, se propuso un modelo de calidad de datos.

El resultado de los estudios realizados, fue la identificación de dieciséis atributos importantes de calidad de datos, basado en la perspectiva de los científicos de datos experimentados, los atributos seleccionados, se pueden observar en la Figura 2.3. Con estos atributos de calidad de datos, los científicos de datos pueden monitorear y mejorar la calidad de los datos para el aprendizaje automático y también establecer los efectos de estos datos en los sistemas de ML.

El siguiente estudio [1], evalúa las métricas de sistemas de ML propuestas en el tutorial: *"Descripción general e importancia de la calidad de los datos para las tareas de aprendizaje automático"*:

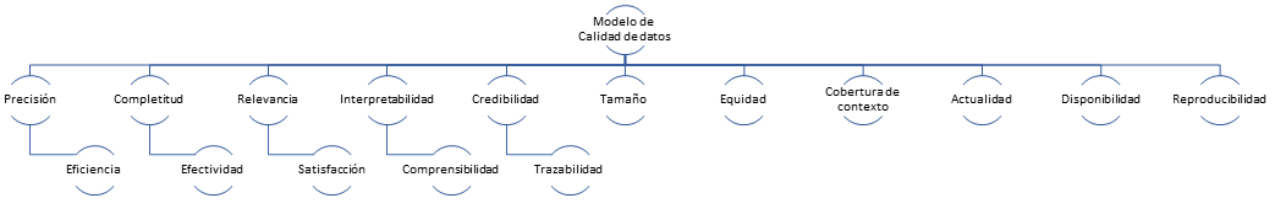


Figura 2.3: Modelo de calidad de datos para ML. Fuente [59].

- Limpieza de datos: Es necesario realizar una limpieza de datos del conjunto de datos seleccionado según la necesidad del sistema de ML.
- Desequilibrio de clases: Distribución desigual de clases en un corpus.
- Ruido de etiquetas: Cuando una parte del conjunto de datos, se encuentra mal etiquetado o tiene errores en sus etiquetas.
- Valoración de los datos: Existen datos dentro del conjunto de datos que son más relevantes para el sistema de ML obteniendo como resultado dos grupos, el primero, con los datos mayor valorados y el segundo con los datos menos valorados.
- Homogeneidad de datos: Se llaman datos homogéneos si todas las entradas siguen un único patrón.
- Transformaciones de datos: Que se le permita al usuario final de los datos, transformar los datos heterogéneos en formatos del usuario homogéneos, mostrando algunos ejemplos de la salida esperada para las entradas dadas.

[66], aborda el problema de calidad de los corpus sobre sistemas críticos basados en ML, analizando algunos estándares de calidad, como **ISO9000**, que mide el grado en que un conjunto de características cumple con los requisitos y **DO-200B**, mide el grado/nivel de confianza de los datos en cuanto al cumplimiento con los requisitos del usuario. En el estudio, se tiene como premisa que los corpus son las principales entradas para los algoritmos de ML, por lo que el cumplimiento de sus funciones depende en gran medida de la calidad de los mismos.

Se propone, un flujo de trabajo para evaluar la calidad de los datos en un contexto

de desarrollo de un sistema certificado basado en ML, organizado en los siguientes artefactos, los cuales son realizados por varios actores:

- Definición estándar del conjunto de datos (DDS): Brindan recomendaciones generales sobre la construcción de los corpus.
- Especificación de requisitos del conjunto de datos (DRS): Recolectan requerimientos aplicables al corpus.
- Plan de verificación del conjunto de datos (DVP): Se definen procedimientos, para verificar el cumplimiento del corpus con sus especificaciones.

La construcción y verificación de corpus requiere la participación de tres expertos diferentes:

- Experto en aplicaciones: Es quien aporta conocimiento en el caso de uso en cuestión, asegura representatividad del corpus y cumple un papel fundamental para el DRS.
- Experto en sistemas de adquisición: Es quien aporta conocimientos sobre parámetros importantes, se anticipa a consecuencias de limitaciones operativas y cumple un papel fundamental para el DRS y DVP.
- Experto en ML: Garantiza buenas prácticas en la creación de corpus, por ejemplo: garantizar la calidad estadística del corpus, el manejo correcto de sesgos, el tamaño del corpus y la coherencia en la anotación. Es quien aporta conocimiento en el DDS y lo adapta para la construcción del DRS.

En conclusión se busca la creación y gestión adecuadas de corpus con condiciones necesarias para confiar en sistemas basados en ML.

Capítulo 3

Estado del arte

En este capítulo, se realizará, una revisión de los corpus orales de emociones. Iniciando, con los corpus orales diseñados para los sistemas de reconocimiento automático del habla, el cuál fue el antecesor de los sistemas de reconocimiento automático de emociones a partir del habla. De allí se revisará, la clasificación de los corpus orales de emociones y finalmente se analizará el diseño y construcción de los corpus orales de emociones desde sus inicios, en 1996 a la actualidad 2021.

3.1. Corpus orales para el reconocimiento automático del habla

El proyecto *Common Voice* [8] se creó debido a la falta de disponibilidad de conjuntos de datos para el entrenamiento y validación en el campo del reconocimiento del habla, inicialmente fue propuesto para el idioma inglés pero posteriormente se vio la necesidad de contar con un corpus que abarque una mayor cantidad de idiomas.

El corpus tiene como objetivo obtener una colección de datos multilingüe de voz transcritos, destinado a la investigación y desarrollo de tecnología del habla.

Para lograr escala y sostenibilidad, el proyecto emplea el crowdsourcing tanto para la recopilación como para la validación de datos.

Hasta donde se sabe, este es el mayor corpus de audio de dominio público para

el reconocimiento de voz, tanto en términos de número de horas como de número de idiomas. Para la mayoría de estos idiomas, estos son los primeros resultados publicados sobre el reconocimiento de voz automático de extremo a extremo.

El corpus Common Voice, es un corpus de habla multilingüe de fuentes múltiples que puede escalar a cualquier idioma a través del esfuerzo de la comunidad. Todos los datos de voz se publican bajo una licencia *Creative Commons*, lo que convierte a Common Voice en el corpus de dominio público más grande diseñado para el reconocimiento automático de voz.

Resumen colección del corpus Common Voice	
Idiomas	38
Señal	Voz
Contenido Corpus	
Número de participantes	50'000
Número de horas	2'500

Cuadro 3.1: Resumen de los resultados de la recopilación de datos y del corpus **Common Voice** [8].

En los últimos años se han producido importantes avances en el área del reconocimiento automático del habla, particularmente para el idioma portugués brasileño, sin embargo los recursos existentes se componen de audios que contienen solo discursos leídos y preparados, dejando un vacío para los corpus de habla espontánea, que son esenciales para las aplicaciones de reconocimiento automático de voz.

CORAA (Corpus de Audios Anotados) [39] es un conjunto de datos disponible públicamente para las aplicaciones de reconocimiento automático del habla que contienen el audio y la transcripción, además posee portugués europeo. El corpus CORAA se ensambló para mejorar los modelos de reconocimiento automático del habla con fenómenos del habla espontánea y así motivar a los jóvenes investigadores a iniciarse en el estudio de reconocimiento automático del habla en portugués.

El corpus CORAA cuenta con 29'077 horas de pares validados (audio-transcripción) compuesto por corpus públicos en portugués brasileño y charlas TEDx en Portugués brasileño y europeo, se puso a disposición del público este nuevo y amplio corpus

para el entrenamiento de modelos de reconocimiento del habla, cerrando la brecha de conjuntos de datos anteriores, con la falta de habla espontánea e informal utilizando conversaciones, diálogos y entrevistas de las fuentes anteriormente mencionadas.

Resumen colección del corpus CORAA	
Idiomas	Portugués brasileño y europeo
Señal	Audio y Transcripción
Contenido Corpus	
Total de horas	29'077
Número de horas portugués europeo	4.69

Cuadro 3.2: Resumen de los resultados de la recopilación de datos y del corpus **CORAA** [39].

Las principales contribuciones realizadas en este trabajo se resumen en la creación de un gran corpus en idioma portugués brasileño de pares validados (audio/transcripción) compuesto por cinco corpus públicos, adaptado para la tarea de reconocimiento automático del habla y se incluyó también portugués europeo. Según el alcance de la investigación es el primer corpus espontáneo del habla en portugués brasileño y está disponible para ser utilizado libremente.

Una de las aplicaciones de los sistemas de reconocimiento automático del habla es la traducción de idiomas, la cuál ha sido testigo recientemente de un resurgimiento de popularidad. Los corpus existentes involucran el idioma Inglés como idioma origen, dominios específicos o pocos recursos, por ello se presenta el corpus CoVost [87], un corpus de traducción de voz a texto multilingüe, del francés, alemán, holandés, ruso, español, italiano, turco, persa, sueco, mongol y chino al inglés. El corpus CoVost está diversificado con más de 11'000 hablantes y más de 60 acentos, proporcionando un punto de referencia inicial que incluye el conocimiento adquirido en la elaboración del corpus para los primeros modelos multilingües de extremo a extremo, muchos a uno, para la traducción de idiomas hablados. El corpus CoVost se publicó con la licencia *Creative Commons* y es de libre uso, así como las muestras de evaluación Tatoeba [1]

[1]Tatoeba (TT) es un corpus de aprendizaje de idiomas creado por la comunidad, que tiene oraciones alineadas en varios idiomas con el habla correspondiente parcialmente disponible.

adicionales utilizadas.

Resumen colección del corpus CoVost	
Idiomas	11
Señal	Audio y Transcripción
Contenido Corpus	
Total de horas	Aproximadamente 629 horas en 11 idiomas.
Participantes	11'000

Cuadro 3.3: Resumen de los resultados de la recopilación de datos y del corpus **CoVost** [87].

Se creó un corpus multilingüe CMU Wilderness [14], el cuál está basado en lecturas grabadas del Nuevo Testamento, proporcionando datos para construir modelos de reconocimiento automático de voz y texto a voz, para potencialmente 700 idiomas. Sin embargo el hecho de que el contenido fuente (la Biblia) sea el mismo para todos los idiomas no se explota hasta la fecha.

Por lo tanto, este corpus, propuso agregar enlaces multilingües, entre segmentos de voz en diferentes idiomas y compartir un gran corpus limpio de 8'130 expresiones, habladas en paralelo, en 8 idiomas.

El corpus construido se le denominó MaSS [95] (Corpus multilingüe de expresiones orales alineadas con oraciones). Los idiomas cubiertos, euskera, inglés, finaldés, francés, hungaro, rumano, ruso y español, permitieron las investigaciones sobre la alineación de voz a voz, así como la traducción de pares de idiomas tipológicamente diferentes.

La calidad del corpus final se atestiguó mediante una evaluación humana realizada en un subconjunto del corpus (100 enunciados, 8 pares de idiomas). Y se mostró la utilidad del producto final en una tarea de recuperación de voz bilingüe.

Resumen colección del corpus MaSS	
Idiomas	8
Señal	Audio y Transcripción
Contenido Corpus	
Total de horas	20 horas de voz en 8 idiomas.
Total de expresiones	8'130

Cuadro 3.4: Resumen de los resultados de la recopilación de datos y del corpus **MaSS** [95].

El corpus LibriSpeech [68] es un gran corpus multilingüe adecuado para la investigación del habla.

El corpus se construyó en base a audio-libros leídos de LibriVox, consta de 8 idiomas, inglés, alemán, holandés, francés, español, italiano, portugés y polaco, que incluyen unas 44'500 horas de inglés y 6'000 horas para los demás idiomas. Además proporcionaron modelos de lenguaje, y referencia de modelos de reconocimiento de voz automático para todos los idiomas.

La generación de un gran corpus transcrito permitirá abrir nuevas vías de investigación de reconocimiento automático del habla y traducción texto a voz. El conjunto de datos está disponible gratuitamente en: http://www.openslr.org.

Resumen colección del corpus LibriSpeech	
Idiomas	8
Señal	Audio y Transcripción
Contenido Corpus	
Total de horas	44.5k horas Inglés y 6k horas otros idiomas.
Total participantes	Aproximadamente 5'980.

Cuadro 3.5: Resumen de los resultados de la recopilación de datos y del corpus **LibriSpeech**. [68]

El corpus TEDx Multilingüe [69] se creó para respaldar la investigación de reconocimiento de voz y traducción de voz en muchos idiomas de origen distintos al inglés.

El corpus se realizó en base a una colección de grabaciones de audio de charlas

TEDx en ocho idiomas de origen. Se segmentaron las transcripciones en oraciones y se alinearon con el audio de idioma de origen y las traducciones del idioma de destino.

El corpus se ha publicado junto con una fuente de código abierta con una licencia CC BY-NC-ND 4.0, que permite la extensión a nuevos discursos e idiomas a medida que estén disponibles. Esta metodología de creación de corpus se puede aplicar a más idiomas y crear conjuntos de evaluación paralela, proporcionando líneas base en múltiples sistemas de reconocimiento automático del habla y traducción de voz, de modo que se pueda mejorar el rendimiento de la traducción para pares de idiomas de bajos recursos.

Resumen colección del corpus TEDx Multilingüe	
Idiomas	100^{+} Lenguajes
Señal	Audio y Transcripción
Contenido Corpus	
Total de grabaciones	150'000*

Cuadro 3.6: Resumen de los resultados de la recopilación de datos y del corpus **TEDx Multilingüe** [69].

(*) Cada año se añaden aproximadamente 3'000 grabaciones.

Los corpus orales revisados en esta sección hacen referencia al reconocimiento automático del habla, lo cuál nos permite conocer cuál es el estado en que se encuentra dicho estudio que es el predecesor del reconocimiento de emociones a partir del habla. El corpus oral que se ha diseñado y construido en esta tesis permite analizar las emociones del hablante mediante la libertad de expresión, lo cuál requiere actividad cognitiva además de la actividad motora.

3.2. Clasificación de los corpus orales de emociones

Según [61, 84, 11] los corpus orales de emociones pueden ser clasificados de la siguiente manera:

3.2.1. Bases de datos del habla de emociones basadas en actores

Las bases de datos del habla de emociones basadas en actores están compuestas por artistas teatrales que son conocedores y competentes, a estos artistas se les pide que expresen lingüísticamente oraciones neutrales en diversas emociones.

Aproximadamente el 60 % de las bases de datos han sido recopiladas para realizar investigaciones donde se puso a disposición una gama completa de emociones. Se dispuso de varias bases de datos en varios idiomas. El discurso actuado por profesionales es el más confiable para el reconocimiento del habla emocional ya que los profesionales pueden emitir un discurso con emociones que presenten una gran amplitud o fuerza. El inconveniente con una base de datos de este tipo es que se describe como deben interpretarse las emociones en lugar de como se revelan, además según [90] las emociones actuadas tienden a ser más exageradas que las reales.

3.2.2. Bases de datos del habla de emociones obtenidas

Las bases de datos del habla de emociones obtenidas se generan simulando situaciones artificiales de emociones sin el conocimiento del hablante o con el conocimiento del hablante. Este escenario es conocido como *Wizard of Oz* para ayudar al actor o participante a alcanzar la emoción esperada utilizando la interacción con una situación externa, un robot o una computadora.

La ventaja de una base de datos de este tipo es que está más cerca a las bases de datos naturales, mientras que la limitación es que no todas las emociones se encuentran disponibles y si los hablantes saben que están siendo filmados entonces simularan sus emociones.

3.2.3. Bases de datos orales de emociones naturales

En las bases de datos orales de emociones naturales, las emociones son expresadas espontáneamente. Es más realista utilizar datos de voz que se recopilan de situaciones de la vida real.

A veces es difícil reconocer las emociones en los discursos o conversaciones con emociones naturales. Por otro lado existe la posibilidad de que algunas emociones estén au-

sentes y además pueden incluir ruido de fondo. Desafortunadamente existen cuestiones legales y morales que prohíben el uso de las mismas para propósitos de investigación.

La ventaja de una base de datos de este tipo es que presenta las emociones en su estado natural por lo que al utilizarla en un entorno con datos reales tendrá mejores resultados que las anteriores. Las bases de datos naturales se registran generalmente a partir de los siguientes escenarios:

- Entrevistas sin guión en radio/televisión en las que los entrevistadores inducen a los hablantes a vivir una experiencia emocionalmente intensa.
- Entrevistas donde el entrevistado habla sobre situaciones felices y tristes de su vida.
- Audios/Videos tomados de programas de televisión, programas de la actualidad.
- Grabaciones de videos ocultas de situaciones de la vida real, por ejemplo pasajeros en el mostrador de equipaje perdido, sesiones de terapia, conversaciones telefónicas,etc.

3.3. Diseño e implementación de los corpus orales de emociones

En los cuadros 3.7, 3.8, 3.9, se puede observar un resumen de los corpus de emociones a partir del habla con información adicional como el idioma, el número y profesión de los participantes, otras señales fisiológicas registradas simultáneamente, el propósito de la recopilación de datos (reconocimiento del habla emocional o síntesis expresiva), los estados emocionales y el tipo de emociones (naturales, simulados y obtenidos).

A continuación de dichos cuadros se analiza el diseño y construcción de los corpus orales de emociones de tipo natural y los que son en idioma español.

En [4] , se genera un corpus en idioma inglés, contó con 22 participantes, los cuáles eran pacientes ancianos deprimidos, entre ellos 12 hombres y 10 mujeres. Las emociones utilizadas en el corpus fue la depresión y neutral. La construcción y diseño del corpus permitió analizar las medidas acústicas en la voz en la depresión, demostrando así que

los pacientes deprimidos tienden a una menor pronunciación y acentuación correctas. Las medidas acústicas del habla del paciente pudieron proporcionar procedimientos objetivos que permitan ayudar a la evaluación de la depresión. El propósito del corpus fue el reconocimiento de la depresión mediante la medida de la fluidez y prosodia en el habla. El tipo del corpus es natural modo audio.

En [5], se construyó y diseñó un corpus con 40 estudiantes, los cuáles expresaron 5 emociones básicas, ira, miedo, alegría, tristeza y disgusto. El diseño del corpus consistía en solicitar a los participantes recordar un evento emocional, adicionalmente se aplicó evaluaciones fisiológicas para evaluar el contenido emocional.Los sujetos que respondieron dentro del criterio establecido en la evaluación fisiológica representaban dicha emoción. A esto le siguió un análisis informático para determinar un conjunto de criterios que pudieran representar cada emoción. Finalmente el método se validó en un corpus oral de emociones en hebreo, el método obtuvo resultados confiables. El corpus es de tipo natural y modo audio además de señales laringográficas, miograma de rostro, respuesta galvánica de la piel y ritmo de latidos del corazón.

En [7], se diseñó y construyó un corpus en idioma inglés, donde participaron varios participantes y usaron las emociones de molestia, diversión, neutral, frustración y cansancio. se investigó el uso de la prosodia para la detección de frustración y molestia en el diálogo natural entre humano-computadora, la contribución de la información del modelo de lenguaje y el .estilo"de habla. Los resultados muestran que un modelo prosódico puede predecir si una expresión es neutra frente a "molesto o frustradoçon una precisión a la par con la concordancia entre etiquetadores humanos. El corpus es de tipo natural modo audio.

En [20], se generó un corpus con un total de 50 participantes. Las emociones que emplea son enojo, miedo, alegría, neutral y tristeza. El idioma utilizado es el inglés. La clase del corpus es semi-natural, ya que 10 participantes leyeron "McGilloway-stylepassagesz otros 10 pasajes, versiones escritas de emociones que ocurren naturalmente en la base de datos natural de Belfast. Se utilizó un guión para la construcción del corpus. El material lingüístico utilizado fueron los pasajes de tamaño de un párrafo redactados en primera persona, los guiones abarcan un periodo en el que la emoción varía en intensidad. El modo del corpus es sólo audio.

En [22], Belfast Natural, se generó un corpus con un total de 125 participantes, entre ellos, 31 hombres y 94 mujeres. Las emociones que utiliza el corpus son varias, entre ellas destacan neutral, enojo, tristeza, alegría, preocupación, decepción, miedo, confianza, interés, enamoramiento, aburrimiento, entre otras. El idioma utilizado para la generación del corpus fue el Inglés. El corpus generado se encuentra dentro del grupo de corpus naturales, ya que se utilizaron clips de televisión, fragmentos de entre 10 a 60 segundos de talk shows, programas de la actualidad y entrevistas realizadas por el equipo de investigación, no se utilizaron guiones, sino que fueron discursos interactivos sin guiones. El modo del corpus es audiovisual, cada clip seleccionado muestra el contexto en el que se produce la emoción y su desarrollo a lo largo del tiempo.

En [27], se generó un corpus con un total de 4 participantes, los cuáles fueron conductores. Las emociones utilizadas fueron neutral y estrés y el idioma del corpus es el inglés. El corpus generado es de tipo natural, ya que el método de construcción se baso en el estudio del habla bajo estrés, utilizando el método de grabación del habla producida en el contexto de conducción a velocidades variables mientras se realizan tareas mentales de carga cognitiva variable. Se eligió el escenario de conducción hablando por teléfono. El modo del corpus es de sólo audio.

En [30], se generó un corpus con un total de 70 pacientes, con diagnóstico de depresión y alto riesgo de suicidio a corto plazo, y 40 individuos normales. Las emociones utilizadas fueron depresión y neutral. El idioma del corpus es el inglés. El corpus generado es de tipo natural, el método para la construcción del corpus fue la realización de grabaciones de audios que permitan analizar y comparar la acústica del habla de muestras separadas de hombres y mujeres compuestas de individuos normales e individuos con diagnósticos de depresión El modo del corpus es sólo audio.

En [32], se generó un corpus con un número desconocido de participantes, las emociones utilizadas fueron depresión y neutral. El idioma empleado fue inglés y español. Se desarrolló, probó y evaluó varios programas bilingües computarizados de reconocimiento de voz para la detección de la depresión que entrevistaron verbalmente a hablantes de inglés y español utilizando la escala del Centro de Estudios Epidemiológicos - Depresión. Los estudios proporcionaron evidencia de que las aplicaciones interactivas bilingües de reconocimiento de voz interactivo fueron generalmente factibles de admi-

nistrar, confiables, válidas y equivalentes a los métodos de entrevista estándar (cara a cara y papel y lápiz). El modo del corpus fue sólo audio.

En [86] se diseñó y construyó un corpus con 32 participantes, 13 mujeres y 19 hombres. Las emociones utilizadas fueron enojo, estrés y tareas generadores de estrés. El idioma empleado en el corpus fue el inglés. El método para la construcción del corpus fue someter a los participantes a situaciones de estrés simulado y real para ello se utilizaron 5 dominios, el primero fueron los estilos del habla (lento, rápido, suave, fuerte, molesto, claro y pregunta), el segundo voz producida en ruido, el tercero tarea de seguimiento de computadora de seguimiento dual, el cuarto fueron las tareas de miedo al movimiento del paciente y quinto datos de análisis psiquiátrico. Por ello el corpus se clasifica como tipo natural y simulado. Se generaron 16'000 enunciados de palabras aisladas, los cuáles se guardaron como audios y archivos transcritos.

En [70], se diseñó y construyó un corpus con 8 actores, 4 hombres y 4 mujeres, los cuáles simularon las 7 emociones básicas, cómo, alegría, tristeza, miedo y sorpresa. El idioma empleado en el corpus es el español. El corpus se construyó mediante la interpretación de dos textos, cada uno de los textos fue repetido con 3 grados de intensidad emocional, lo que permitió la grabación de 336 discursos. El corpus es clasificado de tipo simulado. En este trabajo se describió la metodología utilizada para la validación de los resultados obtenidos en un estudio sobre modelado acústico de la expresión emocional en español. Se obtuvo un conjunto de reglas que describieron el comportamiento de los parámetros más significativos del habla relacionados con la expresión emocional. La validación de los resultados de el estudio se logró mediante el uso de un habla sintética que se ha generado siguiendo las diferentes reglas que obtuvieron para cada emoción.

En [45], se diseñó y construyó un corpus con una cantidad desconocida de participantes en idioma inglés.El enfoque del estudio está en un estudio de caso de detección de emociones negativas y no negativas. El enfoque específico está en un estudio de caso de detección de emociones negativas y no negativas utilizando datos de lenguaje hablado obtenidos de una aplicación de centro de llamadas. En este artículo, se utiliza una combinación de tres fuentes de información (acústica, léxica y discursiva) para el reconocimiento de emociones. Se utilizaron datos del lenguaje hablado obtenidos de una aplicación de centro de llamadas por lo que es de tipo natural y modo audio.

En [56], se diseñó y construyó un corpus con 1 actor, en idioma español. Las emociones utilizadas en el corpus fueron alegría, enojo, tristeza, sorpresa y neutral. Se realizaron dos sesiones donde un actor profesional fue grabado simulando 4 emociones, alegría, tristeza, enojo y sorpresa, además del estado neutral, mediante la lectura de 3 pasajes y 15 oraciones. Las grabaciones fonéticamente y prosodicamente etiquetadas manualmente. Se obtuvieron más de 2000 fonemas por emoción disponibles para el análisis. El tipo del corpus es simulado.

En [65], se diseñó y construyó un corpus con 30 participantes, en idioma inglés, las emociones utilizadas fueron normal, alegría, tristeza, enojo y miedo. El artículo se dividió en dos estudios, el primero elaboró un corpus de 700 expresiones cortas que expresan cinco emociones: felicidad, ira, tristeza, miedo y estado normal (no emocional), que fueron interpretadas por treinta actores no profesionales. El segundo estudio utiliza un corpus de 56 mensajes telefónicos de duración variable (de 15 a 90 segundos) que expresan emociones en su mayoría normales y de enfado que fueron grabados por dieciocho actores no profesionales. El tipo del corpus es simulado y natural y de modo audio.

En [72], se diseñó un corpus con 100 participantes, 25 hablantes nativos de alemán, 16 hablantes nativos de inglés y 59 hablantes nativos de francés. Las emociones utilizadas fueron el estrés y las tareas bajo estrés. Se les pide a los participantes que realicen una prueba de razonamiento lógico en dos condiciones diferentes: 1) centrándose en la tarea sin ninguna perturbación, y 2) trabajando en la tarea mientras atienden una tarea de monitoreo auditivo al mismo tiempo. Cada participante produjo alrededor de 100 oraciones de lectura y varios pasajes de habla espontánea. El corpus es de tipo natural y modo audio.

En [73], SmartKom, se diseñó y contruyó un corpus con 45 participantes, en idioma Alemán. Las emociones utilizadas fueron enojo, insatisfacción y neutral. El corpus SmartKom es el primero de una nueva generación de recursos lingüísticos (LR) diseñado para una recopilación de datos más o menos completa de la comunicación hombre-máquina combinando las modalidades de entrada y salida acústica, visual y táctil. Los participantes son grabados en sesiones de 4.5 minutos de duración mientras que interactúan con el sistema SmartKom. El corpus es de tipo natural y modo audio y

video.

En [77], se diseñó y construyó un corpus en inglés, con 12 participantes, los cuáles eran 12 padres, 6 madres y 6 padres. Las emociones utilizadas fueron aprobación, enojo y prohibición. Se recolectó dos tipos de datos experimentales. En el experimento principal, se recolectó datos acústicos de padres hablando con sus bebés. En el segundo experimento, diferentes oyentes adultos juzgaron si cada expresión se clasificaba mejor como aprobación, oferta de atención o prohibición, y juzgaron la fuerza del mensaje. Se grabaron a los 12 padres con sus bebés de 10 a 18 meses en un cuarto silencioso, cada grabación duró 1 hora, donde los padres jugaron e interactuaron con sus bebés, se les pidió a los padres que utilizaran su voz para mantener a los bebés alejados del peligro. Para cada par de padres e hijos, se seleccionó de 30 a 50 enunciados, cada uno de los cuales constaba de una frase u oración. El corpus es de tipo natural y modo audio.

En [81] se diseñó y construyó un corpus con una cantidad desconocida de participantes, en idioma inglés y utilizó un amplio rango de emociones. Se creo un corpus único de discurso emocional natural que permita la búsqueda automática de ejemplos de tipos muy diferentes de expresión emocional. El corpus consta de aproximadamente 5 horas de muestras de discurso emocional tomadas principalmente de programas de radio y televisión del Reino Unido. El corpus es de tipo natural y obtenido y el modo es sólo audio.

Según [15] el corpus IEMOCAP diseñado en su investigación hizo uso de diez actores los cuales participaron en cinco interacciones diádicas. Los requerimientos que utilizaron para la construcción del corpus IEMOCAP fueron los siguientes:

El corpus debía constar de emociones genuinas, en lugar de monólogos y emociones aisladas el corpus debe tener diálogos naturales, para lograr ello se hizo uso de sesiones con scripts y sesiones espontáneas para obtener las emociones de alegría, enojo, tristeza y frustración, otras emociones también fueron obtenidas según lo dictado por los diálogos entre los actores, Se grabó a muchos actores experimentados.

Estas técnicas están arraigadas en teorías y métodos bien establecidos del teatro, proporcionando manifestaciones emocionales más cercanas a las interacciones naturales. El registro del corpus fue controlado en términos de contenido emocional y lingüístico. El corpus además de contener audio también presentó contenido visual para capturar

información no verbal. Las etiquetas de emociones se asignaron en cuanto a evaluaciones subjetivas por tres personas, los dos participantes del diálogo y un tercero.

El corpus tiene aproximadamente doce horas de datos, los cuáles fueron manualmente segmentados, transcritos y emocionalmente etiquetados con etiquetas categóricas (3 evaluadores) y basadas en atributos (2 evaluadores). Para el nivel de atributos el corpus tiene etiquetas de excitación (calma versus activo), valencia (negativo versus positivo) y dominancia (débil versus fuerte), utilizando una escala *Likert* de cinco puntos. En este estudio se consideran los segmentos en que los tres evaluadores independientes alcanzaron un acuerdo por mayoría de votos en las etiquetas categóricas. Se omitieron los segmentos con habla superpuesta, dando como resultado 4784 turnos del habla.

Resumen colección del corpus IEMOCAP	
Datos	12h
Segmentos	7433
Señal	Voz y Video (Head, face and hands)
Corpus Content	
Número participantes	10 actors
Género	5 mujeres y 5 hombres
Emociones	Alegría, tristeza, enojo y frustración
Dimensiones emocionales	Valencia, excitación y dominancia

Cuadro 3.10: Resumen de los resultados de la recopilación de datos y del corpus IEMOCAP.

Según [2] se propone un corpus multimodal de la vida real (noticias, conversaciones) basado en videoclips grabados de entrevistas de canales de televisión francesa: "EMOTV1".

Para estudiar la influencia de las modalidades de percepción de emociones dos anotadores realizaron la anotación con tres condiciones:

1. Audio sin video.
2. Video sin audio
3. Video con audio

Se siguieron los siguientes criterios para la selección de videos para el corpus EMOTV1:

Se seleccionaron entrevistas de televisión (monólogos), que estuvieran basados en situaciones reales, que tuvieran presencia de emociones y visibilidad de los rostros del hablante y cuerpo superior.Se utilizaron señales multimodales (habla, cabeza, rostro, gaze, gestos, torso) y la lengua utilizada fue francés. Se obtuvieron 14 etiquetas: ira, desesperación, disgusto, duda, exaltación, miedo, irritación, alegría, neutral, dolor, tristeza, serenidad, sorpresa y preocupación. Los anotadores también utilizaron dos dimensiones de valoración clásicas: intensidad y valencia.

Resumen colección del corpus EMOTV1	
Número de videos	51
Tamaño total de videos	12m
Segmentos	4s a 43s
Señal	voz y video
Transcripción	2500 palabras
Contenido Corpus	
Número palabras diferentes	800
Emociones categóricas	14
Dimensiones	Intensidad y Valencia
Número de hablantes	48
Temas	24

Cuadro 3.11: Resumen de los resultados de la recopilación de datos y del corpus EMOTV1.

El corpus elaborado presentó las siguientes ventajas: Comportamientos emocionales naturales, una gran variedad de contextos y anotación a distintos niveles.

Las debilidades del corpus son: Falta de visión en las expresiones faciales, gestos y baja calidad del video.

Encontraron ambigüedad en el etiquetado de las emociones, debido a que no existen patrones básicos para definir cada una de las emociones, por lo que propusieron la siguiente tipología de emociones básicas:

- Emociones de baja intensidad: Cuando el evaluador duda entre neutral y la etiqueta dada.
- Emociones combinadas: Dos emociones ocurren al mismo tiempo.
- Emoción actuada enmascarada: Generalmente las personas no muestran visualmente sus emociones reales.
- Secuencia de emociones: Emociones que ocurren en el mismo segmento de audio.
- Conflicto causa-efecto: Por ejemplo, llorar de alegría.
- Ambigüedad emocional: Es difícil decidir entre dos emociones.

Según[18] se generó el corpus: "HEU Emotion"manualmente de películas, series de televisión, y diversos shows, con múltiples lenguajes, esta conformado por diez emociones y tres modalidades, expresión facial, postura corporal y discurso emocional.

Los criterios utilizados para la recolección de datos fueron los siguientes: Realizaron una lista de emociones, las cuáles querían se encuentren en el corpus, luego se seleccionaron tres motores de búsqueda, se almacenaron las URLs de los videos seleccionados en archivos de texto, los videos fueron seleccionados por cinco miembros del equipo y fueron manualmente editados. El proceso de etiquetado se realizó por quince evaluadores.

En cada video seleccionado, el intérprete debe tener una sola expresión, las tomas en el video deben centrarse en el actor que expresa la emoción, un video largo se puede dividir en varias partes sin embargo cada parte presenta una etapa diferente de expresión emocional. Se pueden incluir varios intérpretes en un solo marco pero deben expresar la misma emoción.

A pesar del proceso realizado se presentó una distribución desigual de emociones, se grabó en un entorno natural por lo que contiene ruido, a pesar de ello se alinea a las aplicaciones del mundo real. A la fecha HEU Emotion es el corpus multimodal más grande.

Resumen colección del corpus HEU Emotion	
Número de videoclips	19004
Plataformas de videos	Tumblr, Google, y Giphy
Fuentes de videos	Películas, Series TV y Shows
Señal	Video y Audio
Contenido Corpus	
Emociones	10
Número de hablantes	9951
Modalidad	Expresión facial, postura y habla
Lenguajes	Chino, Koreano, Tailandés e Inglés

Cuadro 3.12: Resumen de los resultados de la recopilación de datos y del corpus HEU Emotion.

Según [26] se presenta el corpus: "LSSED.^{en} idioma inglés, a gran escala para el reconocimiento de emociones de voz que puedan simular la distribución del mundo real.

Las características del corpus LSSED son:

- Ambos géneros se encuentran representados.
- Cada participante será grabado en una o diferentes sesiones de videos en un entorno de laboratorio con una cámara apuntando a ellos.
- Para obtener los diálogos del corpus, el participante es inducido a varias preguntas mientras sus diálogos se asocian a etiquetas de emociones.
- Los diálogos son anotados con la etiqueta de emoción correspondiente, incluyendo enojo, alegría, tristeza, decepción, aburrimiento, disgusto, expectación, miedo, sorpresa, normal y otros.
- Algunos diálogos tienen más de dos emociones, para solucionar este inconveniente cada diálogo tiene información adicional que incluye el género y la edad del participante.

En este trabajo, se presentó LSSED un corpus desafiante en inglés a gran escala para el reconocimiento de emociones del habla que puede simular una distribución real. Señalaron que los algoritmos existentes tienden a adaptarse a corpus a pequeña escala y, por lo tanto, no pueden generalizarse bien en escenas reales.

Resumen colección del corpus LSSED	
Diálogos	147025
Duración diálogos	10min a 20 min
Duración total	200h
Señal	Voz
Contenido Corpus	
Emociones	11
Número de hablantes	820
Género de hablantes	485 mujeres y 335 hombres
Lenguaje	Inglés
Rango de edad	Jóvenes, Adultos y Mayores

Cuadro 3.13: Resumen de los resultados de la recopilación de datos y del corpus LSSED.

El estudio realizado en [58] se centra principalmente en la investigación y el análisis de métodos de construcción del corpus del habla emocional adulta. Este trabajo estudia el estándar de construcción y el estándar de anotación correspondiente del reconocimiento de voz emocional para el entorno natural, y diseña el esquema específico para la evaluación de la efectividad del corpus.

En base a este estudio se diseña un corpus oral de emociones con participantes adultos basado en un entorno abierto. A continuación se explica en líneas generales el diseño estándar del corpus del habla de emociones empleado.

- Primero, la elección de los audios como uno de los vínculos claves para la construcción del corpus.
- Segundo, al construir un corpus de emociones a partir del habla, se debe intentar mantener lo más que se pueda los atributos naturales.

- Tercero, escoger el tema, debido a que la mayoría de la información del corpus está basada en diálogos, se utilizó un dispositivo de grabación el cuál fue ubicado en cualquier lugar del local, desde donde se pudieran oír las conversaciones.

- Cuarto, los archivos grabados se encuentran en formato wav y la frecuencia de muestreo de 16 khz. Finalmente, el formato del archivo resultante es ID, emoción, la emoción anotada es la que tiene el hablante al momento de hablar, género, edad, nombre y hora.

El esquema de anotación de emociones que utilizaron fue el siguiente:

- Primero, cada emoción se divide en cinco escalas, donde 1 es el más débil y 5 el más fuerte, dividiéndose el proceso de anotación en dos etapas.

- La etapa de predeterminación, donde los expertos necesitan discutir y determinar la especificación de las anotaciones basadas en anotaciones independientes de diez grupos del habla. Cuando la mayoría de los expertos tienen el mismo punto de vista, la anotación formal es llevada a cabo.

- La etapa posterior se da después de la anotación, cuando la confianza del experto se actualiza y el resultado es revisado por el algoritmo *"iterative optimal greedy"*.

- Segundo, cuatro emociones se describen por el modelo de espacio tridimensional, valencia, excitación y dominancia, así los expertos necesitan representar los diferentes estados emocionales a través de tres dimensiones, cada dimensión está dividida en cinco niveles.

- Finalmente, debido a que hay más de una emoción expresada en el habla, según estos dos esquemas de anotación, se puede obtener el nivel de expresión de cada emoción en el habla.

El esquema de evaluación que utilizaron fue el siguiente: existen dos formas de extraer rasgos emocionales:

1. Características locales: Consiste en segmentar las señales de voz en tramas y extraer características por cada segmento de voz, por ejemplo, las características del espectrograma se utilizan a menudo.

2. Características globales: Se basa en toda la muestra para calcular el valor estadístico, actualmente se utiliza un modelo de red neuronal basado en tecnología *Deep Learning*.

Para probar la efectividad del corpus actual se extrae el espectrograma y el conjunto de características eGeMAP, los eGeMAPs son conjuntos de funciones comunes que se utilizan en computación de emoción del habla, contiene dieciocho descriptores de bajo nivel como frecuencia, energía/amplitud, espectro, temporal, cepstrum, etc.

En total son trece mil quinientos diálogos y cuatro emociones. Los datos emocionales se encuentran balanceados y el rango de edades oscila entre los dieciocho a treinta años y algunos entre treinta y cuarenta años.

Resumen colección del corpus Adult Emotional Speech	
Diálogos	13500
Señal	Voz
Contenido Corpus	
Emociones	Alegría, enojo, tristeza y neutral
Dimensiones emocionales	Valencia, excitación y dominancia
Número de hablantes	60
Género de hablantes	30 mujeres y 30 hombres
Rango de edad	Mayoría 18 a 30 años, Minoría 30 a 40 años

Cuadro 3.14: Resumen de los resultados de la recopilación de datos y del corpus **Adult Emotional Speech**.

Según [48] se crea el corpus oral de emociones **MSP-PODCAST CORPUS** basado en grabaciones reales obtenidas de audios compartidos en la web, se combinan algoritmos de ML para recuperar grabaciones con contenido emocional balanceado y un proceso de anotación utilizando *Crowdsourcing*. El corpus posee emociones naturales, contenido emocional balanceado y reducción de costos y labor manual.

Los criterios utilizados para la construcción del corpus fueron: Primero seleccionar y descargar *Podcasts* con contenido emocional balanceado. Los podcasts seleccionados deben contener conversaciones entre diversas personas sobre diversos temas. Las

grabaciones son descargadas bajo licencias Creative Commons (CC).

El proceso de segmentación utilizado fue:

1. Convertirlos con el software Sound eXchange a modo mono canal.
2. Utilizar una tasa de muestreo 16 khz y 16 bit PCM.
3. Seleccionaron *podcasts* con una duración entre tres minutos a ciento noventa minutos.
4. Los *podcasts* seleccionados, tienen uno o más hablantes.
5. Utilizaron una herramienta de diarización para la segmentación.
6. Manualmente segmentaron ciento cinco *podcasts*.
7. Se consideraron segmentos con una duración entre 2.75 a 11 segundos.
8. Eliminaron el ruido de los audios.
9. La evaluación de las anotaciones la realizaron a través de *crowdsourcing Amazon Mechanical Turk*.
10. Se etiquetó con características emocionales (excitación, valencia, dominancia) y emociones categóricas (alegría, tristeza, enojo, sorpresa, miedo, disgusto, desprecio y neutral).
11. Los evaluadores sólo pueden seleccionar una opción, después ellos realizaron anotaciones secundarias donde si pudieron seleccionar más de una opción como: tristeza+frustración.

Con el enfoque presentado en el desarrollo del corpus se incrementó su tamaño a más de veinte y siete horas de datos emocionales, con 18238 segmentos, incrementando a 920 el número de *podcasts* analizados. La distribución del corpus a través de las emociones categóricas no es tan equilibrada como las puntuaciones de las dimensiones emocionales, donde se tienen pocos segmentos para ciertas clases emocionales (por ejemplo: miedo, tristeza).

Resumen colección del MSP-PODCAST CORPUS	
Podcasts	920
Duración podcasts	3m a 190m
Segmentos	2.75s a 11s
Señal	Voz
Contenido Corpus	
Número segmentos	18238
Duración	27h 42m
Emociones	8
Atributos emocionales	Excitación, valencia y dominancia
Número de hablantes	151

Cuadro 3.15: Resumen de los resultados de la recopilación de datos y del MSP-PODCAST CORPUS.

El ser humano se comunica usando una estructura altamente compleja de señales multimodales, empleamos tres modalidades de manera coordinada para transmitir nuestras intenciones: modalidad de lenguaje (palabras, frases y oraciones) modalidad de visión (gestos y expresiones) y modalidad acústica (paralingüística y cambios en los tonos vocales). Comprender esta comunicación multimodal es natural para nosotros sin embargo para los sistemas de inteligencia artificial es una tarea compleja.

El corpus CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity) [12] es el corpus más grande de análisis de sentimiento multimodal y reconocimiento de emociones hasta la fecha.

Para la construcción del conjunto de datos se trabajó con videos de pronunciación de oraciones de más de 1000 hablantes en línea de YouTube, todas las oraciones pronunciadas se eligieron aleatoriamente, de varios temas y videos de monólogos. Los videos se seleccionaron, se transcribieron y fueron debidamente puntuados.

El conjunto de datos contiene más de 23'500 videos de pronunciación de oraciones, presenta equilibrio de género, una diversidad de 250 temas y en cuánto a la distribución emocional y de sentimientos no es homogénea, presenta una mayor representatividad entre los sentimientos positivos así como para las emociones.

Resumen colección del corpus CMU-MOSEI	
Idiomas	Inglés
Señal	Lenguaje, Visual y Acústico
Número de videos	3228
Número de temas	250
Contenido Corpus	
Total de horas	65:53:36
Total oraciones	23'453
Participantes	1000

Cuadro 3.16: Resumen de los resultados de la recopilación de datos y del corpus **CMU-MOSEI**.

En este trabajo se modeló un corpus de análisis de sentimientos y reconocimiento de emociones multimodal, el corpus amplía los horizontes de los estudios del lenguaje multimodal humano en el procesamiento de lenguaje natural.

3.4. Discusión

Luego, del análisis realizado, de los corpus orales de emociones, resalta el uso de una gran diversidad de participantes para el diseño y construcción de los diversos corpus, como pacientes, estudiantes, nativos del lugar, de programas de televisión, conductores, actores y pasajeros, dicha variedad permitió conocer las diversas formas de obtención natural de emociones, tanto positivas como negativas. También se registró señales fisiológicas sincrónicas adicionales, como la indicación de sudor, la frecuencia cardíaca, la presión arterial y la respiración, las cuales proporcionan, una verdad fundamental, para el grado de excitación o estrés de los participantes. Se encontró, una evidencia, entre las señales fisiológicas y el grado de excitación de las emociones, más no en el carácter positivo o negativo de las mismas.

Sin embargo en cuánto al idioma empleado en dichos corpus, predominó el inglés y unos cuantos en alemán, hebreo y español. Además, se puede observar la limitación en cuanto a la cantidad de emociones, los cuales abarcan entre 5 o 6 emociones.

En cuanto, a los corpus naturales construidos en base al material de radio o televisión, este siempre está disponible, aunque dicho material plantea el factor de derecho de autor, el cual, impide la distribución de la recopilación de datos, por lo que hay que considerar solo grabaciones que se encuentren libres de derechos, como por ejemplo, entrevistas con especialistas, diálogos en situaciones de la vida real, padres que protegen a sus hijos de situaciones peligrosas, entrevistas entre médicos y pacientes, así como grabaciones entre participantes y máquinas que generan situaciones de estrés o calma.

Referencia	Lenguaje	Hablantes	Emociones	Tipo
Abelin and Allwood (2000)	Sueco	1 Nativo	Ar, Fr, Jy, Sd, Se, Dt, Dom, Sy	Simulada
Alpert et al. (2001)	Inglés	22 pacientes, 19 saludables	Dn, Nl	Natural
Alter et al. (2000)	Alemán	1 mujer	Ar, Hs, Nl	Simulado
Ambrus (2000), Interface	Inglés, Esloveno	8 actores	Ar, Dt, Fr, Nl, Se	Simulado
Amir et al. (2000)	Hebreo	40 estudiantes	Ar, Dt, Fr, Jy, Sd	Natural
Ang et al. (2002)	Inglés	Muchos	An, At, Nl, Fd, Td	Natural
Banse and Scherer (1996)	Alemán	12 actores	H/C, Ar, Hs, Sd,...	Simulado
Batliner et al. (2004)	Alemán, Inglés	51 niños	Ar, Bd, Jy, Se	Obtenidas
Bulut et al. (2002)	Inglés	1 actriz	Ar,Hs, Nl, Sd	Simulado
Burkhardt and Sendlmeier (2000)	Alemán	10 actores	Ar, Fr, Jy, Nl, Sd, Bm, Dt	Simulado
Caldognetto et al. (2004)	Italiano	1 nativo	Ar, Dt, Fr, Jy, Sd, Se	Simulado
Choukri (2003), Groningen	Holandés	238 nativos	Desconocido	Simulado
Chuang and Wu (2002)	Chino	2 actores	Ar, Ay, Hs, Fr, Se, Sd	Simulado
Clavel et al. (2004)	Inglés	18 de TV	Nl, niveles de Fr	Simulado
Cole (2005), Kids' Speech	Inglés	780 niños	Desconocido	Natural
Cowie and Douglas-Cowie (1996),Belfast Structured	Inglés	40 nativos	Ar, Fr, Hs, Nl, Sd	Natural
Douglas-Cowie et al. (2003), BelfastNatural	Inglés	125 de TV	Varios	Semi-Natural
Edgington (1997)	Inglés	1 actor	Ar, Bm, Fr, Hs, Nl, Sd	Simulado
Engberg and Hansen (1996), DES	Danés	4 actores	Ar, Hs, Nl, Sd, Se	Simulado
Fernandez and Picard (2003)	Inglés	4 conductores	Nl,Ss	Natural
Fischer (1999), Verbmobil	Alemán	58 nativos	Ar, Dn, Nl	Natural
France et al. (2000)	Inglés	70 pacientes, 40 saludables	Dn,NI	Obtenido
Gonzalez (1999)	Inglés, Español	Desconocido	Dn, Nl	Obtenido
Hansen (1996), SUSAS	Inglés	32 varios	Ar,Ld, eff., Ss, TI	Natural, simulado

Cuadro 3.7: Colecciones de datos de emociones a partir del habla (orden alfabético en base a las referencias)[84]

Referencia	Lenguaje	Hablantes	Emociones	Tipo
Hansen (1996), SUSC-0	Inglés	18 No-nativos	Nl,Ss	A-stress
Hansen (1996), SUSC-1	Inglés	20 nativos	Nl, Ss	P-Stress
Hansen (1996), DLP	Inglés	15 nativos	Nl, Ss	C-stress
Hansen (1996), DCIEM	Inglés	Desconocido	NI, Sleep deprive	Obtenido
Heuft et al. (1996)	Alemán	3 nativos	Ar, Fr, Jy, Sd,...	Simulado, Elicitado
Iida et al. (2000), ESC	Japonés	2 nativos	Ar, Jy, Sd	Simulado
Iriondo et al. (2000)	Español	8 actores	Fr, Jy, Sd,Se,...	Simulado
Kawanami et al. (2003)	Japonés	2 actores	Ar, Hs, Nl, Sd	Simulado
Lee and Narayanan (2005)	Inglés	Desconocido	Negativo - positivo	Natural
Liberman (2005), Emotional Prosody	Inglés	Actores	Anxty, H/C, Ar, Hs, NI, Pc, Sd, Se,...	Simulado
Linnankoski et al. (2005)	Inglés	13 nativos	An, Ar, Fr, Sd,...	Obtenido
Lloyd (1999)	Inglés	1 nativo	Estrés fonológico	Simulado
Makarova and Petrushin (2002), RUSSLANA	Ruso	61 Nativos	Ar, Hs, Se, Sd, Fr, NI	Simulado
Martins et al. (1998), BDFALA	Portugués	10 nativos	Ar, Dt, Hs, Iy	Simulado
McMahon et al. (2003), ORESTEIA	Inglés	29 nativos	Ae, Sk, Ss	Obtenido
Montanari et al. (2004)	Inglés	15 niños	Desconocido	Natural
Montero et al. (1999), SES	Español	1 actor	Ar, Dt, Hs, Sd	Simulado
Mozziconacci and Hermes (1997)	Holandés	3 nativos	Ar, Bm, Fr, Jy, Iy, NI, Sd	Simulado
Niimi et al. (2001)	Japonés	1 hombre	Ar,Jy,Sd	Simulado
Nordstrand et al. (2004)	Sueco	1 nativo	Hs, Nl	Simulado
Nwe et al. (2003)	China	12 nativo	Ar, Fr, Dt, Jy,...	Simulado
Pereira (2000)	Inglés	2 actores	H/C Ar, Hs, Nl, Sd	Simulado
Petrushin (1999)	Inglés	30 nativos	Ar, Fr, Hs, Nl, Sd	Simulado
Polzin and Waibel (2000)	Inglés	Desconocido	Ar, Fr, Nl, Sd	Simulado
Polzin and Waibel (1998)	Inglés	5 estudiantes teatro	Ar, Fr, NI, Sd	Simulado
Rahurkar and Hansen (2002), SOQ	Inglés	6 soldados	5 niveles de estrés	Natural
Scherer (2000b) Lost Luggage	Varios	109 pasajeros	Ar, Hr, Ie, Sd, Ss	Natural
Scherer (2000a)	Alemán	4 actores	Ar, Dt, Fr, Jy, Sd	Simulado
Scherer et al. (2002)	Inglés, Alemán	100 nativos	2TI, 2Ss	Natural

Cuadro 3.8: Colecciones de datos de emociones a partir del habla (Continuación) [84]

Referencia	Lenguaje	Hablantes	Emociones	Tipo
Schiel et al. (2002), SmartKom	Alemán	45 nativos	Ar, Dn, Nl	Natural
Schröder and Grice (2003)	Alemán	1 Hombre	Suave, modal,fuerte	Simulado
Schröder(2000)	Alemán	6 nativos	Ar, Bm, Dt, Wy,...	Simulado
Slaney and McRoberts (2003), Babyears	Inglés	12 nativos	Al, An, Pn	Natural
Stibbard (2000), Leeds	Inglés	Desconocido	Amplío rango	Natural, obtenido
Tato (2002), AIBO	Alemán	14 nativos	Ar, Bm, Hs, Nl, Sd	Obtenido
Tolkmitt and Scherer (1986)	Alemán	60 nativos	Ss Cognitivo	Obtenido
Wendt and Scheich (2002), Magdeburger	Alemán	2 actores	Ar, Dt, Fr, Hs, Sd	Simulado
Yildirim et al. (2004)	Inglés	1 actriz	Ar, Hs, Nl,Sd	Simulado
Yu et al. (2001)	Chino	Nativo de TV	Ar, Hs, Nl, Sd	Simulado
Yuan (2002)	Chino	9 nativos	Ar, Fr, Jy, Nl, Sd	Obtenido

Cuadro 3.9: Colecciones de datos de emociones a partir del habla (Continuación) [84]

Abreviaturas por emociones: Las categorías de emociones se abreviaron de la siguiente forma: Ae: Molestia, Al: Aprobación, Ar: Ira, An: Atención, Anxty: Ansiedad, Bm: Aburrimiento, Dn: Insatisfacción, Dt: Disgusto, Fr: Miedo, Hs: Felicidad, Ie: Indiferencia, Iy: Ironía, Jy: Alegría, Nl: Neutral, Pc: Pánico, Pn: Prohibición, Se: Sorpresa, Sd: Tristeza, Sk: Conmoción, Ss: Estrés, Sy: Timidez, Wy: Preocupación. Los puntos suspensivos indican que se registraron emociones adicionales.

Abreviaturas para otras señales: BL: Examen de sangre, BP: Presión arterial, H: Frecuencia cardíaca, IR: Cámara infrarroja, LG: Laringógrafo, R: Respiración, V: Video.

Otras abreviaturas: H/C: frío/calor, Ld eff.: efecto Lombard.

Capítulo 4

Método

En este capítulo, se tratará el pipeline del proceso del diseño y construcción del corpus oral de emociones, el cuál incluye la selección de la plataforma, la selección de audios/videos, el pre-pocesamiento de los audios seleccionados, la segmentación y etiquetado del corpus y finalmente, el esquema de validación del corpus.

A continuación, se muestra, el pipeline del proceso de diseño y construcción del corpus oral en idioma español con acento peruano (ver figura 4.1):

1. Selección de la plataforma de audios/videos, la cuál permitirá realizar la búsqueda de datos así como la extracción de los mismos.

2. Selección de los audios/videos, se contará con criterios de inclusión y exclusión para poder seleccionar los audios/videos más adecuados para el diseño y construcción del corpus.

3. El pre-procesamiento del corpus, el cuál es el proceso de normalización de los audios obtenidos de la plataforma seleccionada para su posterior procesamiento de segmentación y etiquetado.

4. Segmentación de los audios previamente seleccionados y pre-procesados y numeración de cada uno de los segmentos a etiquetar.

5. Anotación de las dimensiones emocionales, una vez se tengan los audios segmentados y numerados se realizará el proceso de etiquetado de cada uno de los segmentos del habla con una dimensión emocional, este proceso será realizado por tres anotadores.

6. Consolidación de las etiquetas, teniendo las etiquetas de cada uno de los segmentos de audio por cada uno de los etiquetadores se llevará a cabo el proceso de consolidación de las mismas.

7. Análisis de la coincidencia del etiquetado mediante la evaluación Alpha Cronbach.

8. Análisis de la diversidad emocional en el corpus.

9. Finalmente la calidad del corpus será evaluada de manera cualitativa y cuantitativa.

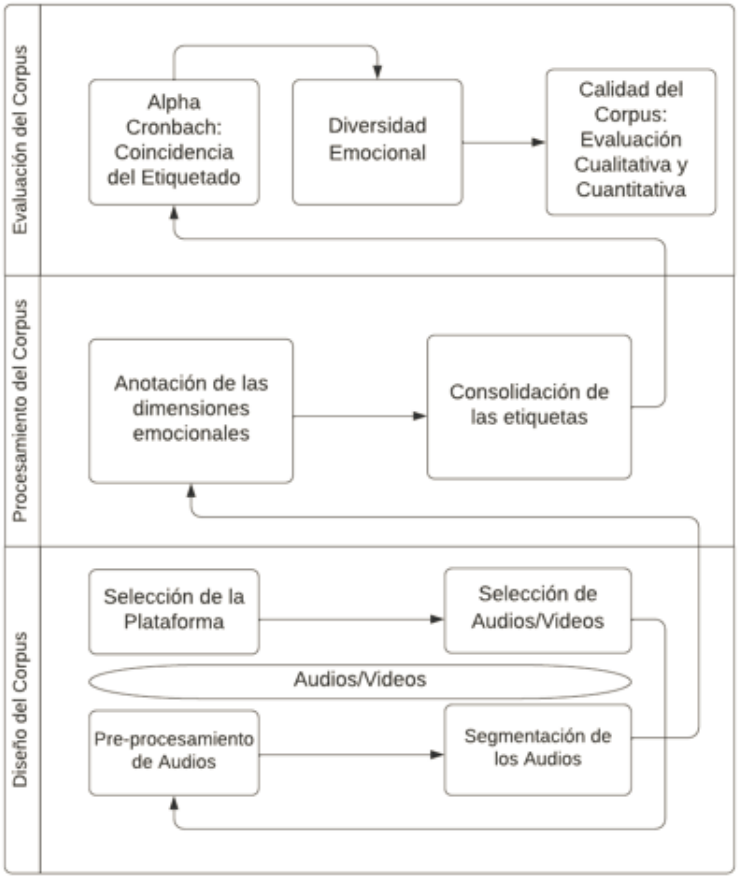


Figura 4.1: Pipeline del diseño y construcción del corpus oral en idioma español con acento peruano. Fuente propia.

4.1. Selección de la plataforma

Actualmente existe una gran diversidad de plataformas en Internet, sin embargo, no todas estas plataformas pudieron ser utilizadas en la generación de un corpus oral. Se establecieron dos criterios para determinar el más adecuado, los cuáles fueron diversidad y accesibilidad.

La diversidad, hace referencia en este trabajo, a las plataformas que brindan una variedad de audios disponibles, en cuánto a temas y oradores y la accesibilidad, se refiere, al hecho de que las plataformas brinden un acceso ilimitado a los audios y las búsquedas de los mismos no impongan restricciones.

Con estos dos criterios se identificaron inicialmente tres plataformas candidatas:

Spotify, *Ivoox* y *YouTube*, la que mejor se adapto a las necesidades fue *YouTube*, ya que cuenta con una mayor diversidad de videos, temas y ponentes y permite una búsqueda más amplia a diferencia de *Spotify* e *Ivoox*.

4.2. Selección de los audios/videos

Una vez que la plataforma es seleccionada los audios candidatos deberán ser identificados aplicando los siguientes criterios.

4.2.1. Criterios de inclusión

Los criterios para incluir un audio fueron los siguientes:

1. Seleccionar audios con contenido natural y espontáneo.
2. Seleccionar audios con licencia CC o que el autor explícitamente haya brindado su consentimiento para el uso de los mismos, de modo que al finalizar la selección de audios, estos, puedan ser descargados libremente, sin restricciones, para generar un corpus disponible al público.
3. Seleccionar audios utilizando palabras claves, que nos permitan conocer preliminarmente si se abordará un tema con emociones positivas o negativas, por ejemplo, entrevistas de temas polémicos, como debates o temas extraños, así como reportajes geográficos del Perú, charlas relacionadas al folclore de nuestra ciudad y testimonios de emociones negativas, como sucesos lamentables ocurridos en nuestro país y testimonios que representen emociones positivas, como sucesos de éxito en nuestro país.
4. Seleccionar audios en español, con los diferentes acentos peruanos, en los tipos de dialecto estándar o vernacular.
5. Mantener el equilibrio entre los géneros de los participantes, incluyendo audios con diferentes hablantes entre hombres y mujeres.

4.2.2. Criterios de exclusión

Los criterios de exclusión de audios, fueron los siguientes:

1. Los audios, con acentos que no pertenezcan a la diferentes ciudades del Perú, no serán seleccionados.

2. Los audios, que al escucharlos en el inicio, intermedio y fin, predomine la emoción neutral, no serán seleccionados.

3. Los audios, pertenecientes a grabaciones actuadas serán excluidos, como películas, series o programas previamente ensayados por actores.

4.3. Pre-procesamiento

Es el procedimiento en el cuál los audios se preparan para la segmentación y etiquetado. La herramienta utilizada para esta actividad fue AUDACITY[1]. El proceso consiste en convertir los audios en un formato más consistente: audios en formato wav, monocanal y formato de 16 bit PCM.

4.4. Segmentación

La segmentación, es el proceso por el cual, el conjunto de audios es recibido como entrada y por cada uno de los audios seleccionados, los segmentos, son generados como la salida. El paso inicial involucra la generación automática de los segmentos utilizando la herramienta PRAAT[2], haciendo uso del script *mark_pauses*, el cual utiliza los silencios presentes en el audio, como criterio de segmentación. El siguiente paso, comprende, por cada segmento automáticamente generado, la verificación de los siguientes criterios de elegibilidad:

- Tener solo la voz de un hablante. Los segmentos que tienen la voz de dos o más hablantes simultáneamente o no son descartados.

[1] https://www.audacityteam.org/
[2] https://www.fon.hum.uva.nl/praat/

- Tener una duración mínima entre tres y quince segundos, aproximadamente. Los segmentos que sean más cortos serán descartados, los segmentos más largos pueden ser divididos en dos segmentos más cortos.

- Tener una frase o palabra que exprese una emoción. Los segmentos donde se percibe más de una emoción serán divididos o descartados.

La verificación del desempeño de estos segmentos se realizó manualmente, los segmentos de entrada fueron escuchados por una sola persona, uno por uno en su totalidad, dicha persona, debe tener conocimiento de los criterios de elección de segmentos, los cuáles han sido propuestos por expertos [49]. El paso final en el proceso de segmentación fue la asignación de un identificador secuencial numérico, a cada segmento que haya pasado el criterio de elegibilidad. La identificación de los segmentos, fue llevada a cabo, para poder realizar las siguientes actividades sin ambigüedad.

4.5. Etiquetado

El etiquetado, es un procedimiento, que recibe como entrada la salida del proceso de segmentación y genera un conjunto de atributos emocionales, un vector (valencia, excitación y dominancia) por cada segmento. Los segmentos de audio y sus etiquetas correspondientes, forman el corpus, compuesto de n muestras etiquetadas $S = z_i{}_{i=1}^{n}$, donde $z_i = (x_i, y_i)$, $x_i \in X$, y $y_i \in R^3$. El etiquetado comprende dos actividades principales: el etiquetado individual y el consenso. En la primera actividad, se consideraron tres etiquetadores, de los cuáles, cada etiquetador, generará un conjunto de anotaciones A_j, individualmente e independientemente. Luego los elementos del conjunto A_j son clasificados en dos grupos: los que coinciden en las tres dimensiones emocionales y los que difieren en al menos una dimensión. Las etiquetas coincidentes pasan a formar parte del corpus final. Para las etiquetas no coincidentes, el valor promedio de los atributos emocionales de cada segmento se calcula para asignar el valor de la etiqueta final. Las anotaciones de los segmentos de audio se realizaron utilizando la herramienta PRAAT, cada segmento fue etiquetado con tres atributos emocionales, de acuerdo a la escala Self Assesment Manikin (SAM) que se muestra en la Figura 4.2.

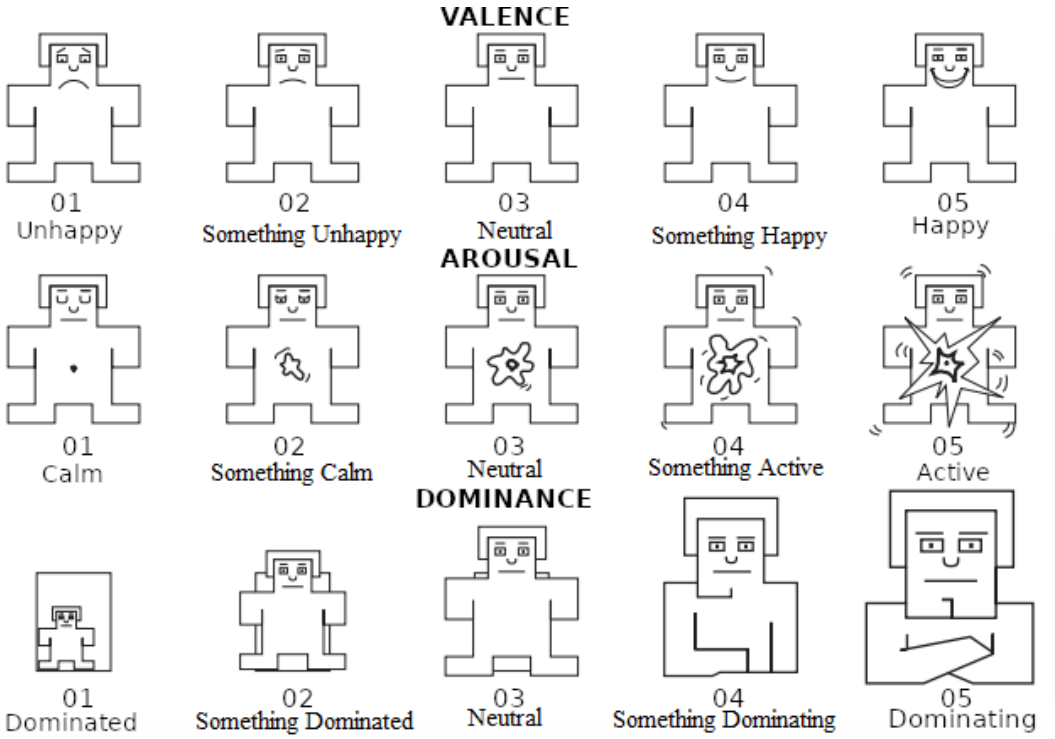


Figura 4.2: La escala de Self-Assessment Manikin. Fuente [48]

La escala SAM, fue seleccionada, debido a que es una de las más utilizadas en los corpus orales de emociones, revisados en el estado del arte [16, 17, 49, 58]. Además que al iniciar con el proceso de etiquetado de las emociones en el corpus, se vió la necesidad de poder contar además de la descripción de la dimensión emocional, con la imagen de esta, ya que al ser un corpus con expresiones genuinas, era necesario poder identificar la emoción del hablante con la imagen mostrada en la escala SAM. De esta manera, la realización del proceso de etiquetado, fue más sencillo.

Finalizando la etapa de etiquetado por cada uno de los etiquetadores se utilizó el script de PRAAT *save_labeled_intervals_to_wav_sound_files* para guardar cada uno de los segmentos en formato WAV con sus respectivas etiquetas. Los atributos emocionales (valores de valencia, excitación y dominancia) fueron colocados como nombre a cada uno de los archivos WAV.

4.6. Esquema de evaluación del corpus oral de emociones

La base de la evaluación de la eficacia real, del corpus oral de emociones, se da, cuando el reconocimiento de emociones se lleve a cabo, con el corpus diseñado y construido en esta tesis, de modo que permita el reconocimiento de emociones en español con acento peruano.

4.6.1. Modelos de deep learning

Red neuronal convolucional

Una red neuronal convolucional *CNN* [88], es un algoritmo, el cuál, reconoce patrones en los datos, generalmente, las redes neuronales, están compuestas por una colección de neuronas, las cuáles, se organizan en capas, cada una con sus propios pesos y sesgos.

Los componentes básicos de una CNN, son:

1. Un tensor, que representa un matriz n-dimensional.
2. Una neurona, que representa una función, la cuál recibe muchas entradas y devuelve una salida.
3. Una capa, que es una colección de neuronas, que realizan, una misma operación.
4. Los pesos y sesgos kernel, son únicos para cada neurona, se ajustan durante la fase de entrenamiento y permiten, que el clasificador, se adapte al problema y al conjunto de datos proporcionado.
5. Una CNN, transmite, una función de puntaje diferenciable, que se representa, como puntajes de clase, en la capa de salida.

Los pasos, que se realizan, en cada capa, de una CNN, son:

1. Capa de entrada: representa a la imagen de entrada en la CNN.
2. Capas convolucionales: es la capa fundamental, ya que contienen los kernels aprendidos (pesos), que extraen características, que distinguen, imágenes entre sí, para producir un mapa de activación.

3. Capas pooling: La capa de pooling resume un área $p \times p$ y se utiliza para reducir el tamaño de la imagen, mediante el *stride*, indica, cuántos píxeles, se debe desplazar el kernel a la vez.

4. Capas Flatten: Esta capa, convierte, una capa tridimensional en la red, en un vector unidimensional, para adaptarse a la capa de entrada. Las capas convolucionales anteriores de la red, extraían las características de la imagen de entrada, pero ahora es el momento de clasificar las características. Se utiliza la función softmax, para clasificar estas características, lo que requiere una entrada unidimensional.

Las funciones, con las que trabaja la CNN son:

1. RELU: La función más ampliamente usada, debido a su baja complejidad computacional.

$$RELU(x) = max(0, x)$$

2. Softmax: El propósito de las operaciones softmax, es asegurarse, que las salidas de la CNN, sumen 1.

Red neuronal residual

Una red neuronal residual *ResNet* [36], es un algoritmo, el cuál, al igual que las demás redes neuronales, están compuestas por una colección de neuronas, las cuáles, se organizan en capas, pero, con la diferencia, que en lugar de contar con una decena de capas, puede contar con centenas de capas.

La propuesta, de este algoritmo de redes neuronales, es la siguiente:

1. Resolver el problema de la reducción del gradiente, en capas anteriores, cuando se tienen muchas capas.

2. Obtener los resultados residuales de un grupo de capas, para utilizarla en la conexión de salto, conocida como el *skipped connection*.

El funcionamiento del aprendizaje residual, es el siguiente:

1. Al tener una red neuronal, con un numero x de capas, que tiene un funcionamiento óptimo.

2. Entonces, se propone añadir más capas, pero, antes de añadir dichas capas, se obtiene el mapeo de salida de las x capas que funcionan de manera óptima, pasando este mapeo, a la capa resultado.

3. Lo que realiza un atajo, el cuál aplica las funciones identidad a la salida.

Las ventajas, de las redes neuronales residuales son:

1. Las funciones residuales, que son utilizadas en las ResNet, al utilizar pesos que tiendan a cero, se convertirán, en funciones identidad, ya que la salida, es igual a la entrada x.

2. Por lo que, al añadir bloques, no modificarán el comportamiento de la red.

4.6.2. Particionamiento del corpus

En los modelos de deep learning, se realiza el particionamiento del corpus, uno para el entrenamiento, otro para la validación y otro para la realización de las pruebas. En el experimento se utilizaron 3749 segmentos de audio válidos. El 80 % de los segmentos (3000 segmentos) se utilizó para la etapa de entrenamiento y el 20 % (749) para la etapa de validación.

4.6.3. Métricas de evaluación

El corpus, diseñado y construido en esta tesis, es un corpus de tipo natural, el cuál ha sido etiquetado, con tres dimensiones emocionales, es por ello, que se denomina, como un corpus multiclase, ya que posee más de una etiqueta en un solo segmento, entonces, las métricas que mejor se ajustan para realizar dicha evaluación, son, la función de error (MSE) y la función de pérdida basada en la correlación con el coeficiente de correlación de concordancia (CCC) [10]. Dichas métricas, se encuentra detalladas en el marco teórico, como estándar de evaluación. Para ello, fue necesario transformar las etiquetas del corpus, que se encontraban en la escala de [1,5] a una escala [-1,1],

debido, a que es más legible, ya que la valencia, la excitación y la dominancia, según sus definiciones, van de un rango negativo a positivo.

4.6.4. Preprocesamiento

Todos los segmentos del corpus, los cuáles son audios, son convertidos a espectogramas, y posteriormente a una representación Coeficientes Cepstrales de Frecuencia Mel (MFCC), que representan, coeficientes, para la representación del habla, basados en la percepción auditiva humana. El conjunto de características MFCC, son alrededor de diez a veinte características, que describen, de manera concisa un sobre espectral.

Debido a que, cada segmento, tiene una longitud diferente, se realizó un preprocesamiento, utilizando una ventana deslizante, para extraer frames de igual longitud, por cada segmento. Los paramétros utilizados fueron: $WindowSize = 55$ y $Hop\_length = 10$.

4.6.5. Entrenamiento

El modelo de la red neuronal CNN, empleado, para realizar el entrenamiento del corpus fue:

```
Entrada: 32 frames
5 Capas Convolucionales
1 Capa Full Connected
```

Listing 4.1: Modelo de red neuronal convolucional

El modelo de la red neuronal residual, empleado, para realizar el entrenamiento del corpus fue:

```
Entrada: 32 frames
17 Capas Convolucionales
1 Capa Full Connected
```

Listing 4.2: Modelo de red neuronal residual

Ambas, recibieron como batch de entrada, 32 frames de audio 2D, de tamaño $20x55$, se utilizó la función de optimización *RELU* y para la etapa de evaluación de los resultados, se utilizó la funciones de coste *MSE* y *CCC*.

La red neuronal convolucional, con la función de coste, CCCL, tomo 3462 segundos de entrenamiento en 6000 iteraciones, aproximadamente. Utilizando la función de coste, MSE, el entrenamiento tomó 3497 segundos en 6000 iteraciones, aproximadamente.

La red neuronal residual, con la función de coste, CCCL, tomo 2163 segundos de entrenamiento en 6000 iteraciones, aproximadamente. Utilizando la función de coste, MSE, el entrenamiento tomó 2179 segundos en 6000 iteraciones, aproximadamente.

El criterio de parada, utilizó la estrategia *Early Stopping* para evitar el *overfitting*, el cuál, permite que el modelo sea generalizable a los distintos datos y no se estanque en un solo modelo de datos.

El resultado, fue el grado de predicción para la valencia, excitación y dominancia de los segmentos.

4.7. Discusión

En base al diseño propuesto en este capítulo, para la construcción del corpus oral de emociones, se diseñó una evaluación cualitativa y cuantitativa, de modo que se pueda validar la eficacia real del corpus, en base a los conceptos empleados en su diseño y construcción y al reconocimiento de emociones.

Capítulo 5

Evaluación

En está sección se realizará la evaluación del corpus diseñado y construido, para lograr ello se ha dividido la evaluación en dos partes, la evaluación cualitativa, la cuál permite analizar las características del corpus desde la perspectiva de los conceptos involucrados en su diseño y construcción. Mientras que la evaluación cuantitativa se orienta hacia los objetivos que se desean estudiar utilizando métodos que permitan cuantificar si estos fueron obtenidos de forma total o parcial.

A continuación se detallará cada uno de los conceptos y métodos utilizados para poder llevar a cabo la evaluación.

5.1. Evaluación cualitativa

5.1.1. Registro de datos

Audio/Video Seleccionados

Con los criterios de selección de audios/videos mencionados en el punto 4.2. fueron seleccionados los audios/videos del Cuadro 5.1:

ID	Nombre	URL	Tiempo	Perfil
1	Debate Presidencial 2021	https://youtu.be/4ejMuYd9SO4	01:28:18	N/P
2	Testimonios de Flor Huilca y Martha Flores	https://youtu.be/ez_4KGuirLE	00:31:00	N
3	Testimonios de la Clasificación de Perú al Mundial 2018	https://youtu.be/BCKtag_eQdw	00:17:11	P
4	Reportaje de tragedia en Arequipa 1996	https://youtu.be/RV9M8pIg6PA	00:12:47	N
5	Reportaje de heladas en Arequipa	https://youtu.be/7aLDKH6Lnp0	00:20:09	N
6	Reportaje de Huaynacotas en Arequipa	https://youtu.be/tTNFWbsCUi8	00:30:59	P
7	Entrevista ataque piloto de la FAP a OVNI en la Joya, Arequipa	https://youtu.be/OTEi6k-4roo	01:18:04	P
8	Testimonio Maria Antonieta Quispe	https://youtu.be/a4TbD2QXAhE	00:27:52	N
9	Reportaje fin de semana en Arequipa	https://youtu.be/RbTpXVzYt10	00:16:09	P
10	Entrevista ”Perú derechos sexuales y reproductivos”	https://youtu.be/EtcZG-7ngPk	00:55:32	N
11	Reportaje muerte Maryori en Huancayo	https://youtu.be/JktSxUEUmJY	00:10:55	N
12	Charla TedxCharacato Monica Huerta	https://youtu.be/2IqJaJM3-lw	00:20:28	P
13	Reportaje de restaurante peruano en México	https://youtu.be/-lRp9B6Z35w	00:21:50	P
14	Testimonio Cipriana Huamani	https://youtu.be/Da0tn2YEKHY	00:34:37	N

Cuadro 5.1: Videos Seleccionados.

Abreviatura por perfil: Las categorías del perfil se abreviaron de la siguiente forma: N: Negativo, P: Positivo.

Resumen del corpus

El corpus contiene siete horas cuarenta y cinco minutos y cincuenta y dos segundos de audio, compuesto por un total de 3749 segmentos de habla natural. El cuadro 5.2 resume los resultados de la compilación de audios y el contenido del corpus.

Resumen colección del corpus	
Número de videos	14
Tamaño total de videos	7h 45m 52s
Número de etiquetadores	3
Señal	voz
Contenido Corpus	
Número segmentos de audio	3749
Longitud total de los segmentos	7h 45m 52s
Dimensiones	Valencia, Excitación y Dominancia
Escala de etiquetas	01, 02, 03, 04, 05
Número de hablantes	80
Género de hablantes	36 mujeres y 44 hombres
Rango edad hablantes	Adultos

Cuadro 5.2: Resumen de los resultados de la recopilación de datos y el conjunto de datos.

Contenidos del corpus

El corpus esta disponible en el repositorio ZENODO en un paquete ZIP. Los 3749 segmentos de audio en formato WAV están organizados en catorce carpetas correspondientes a catorce fuentes de videos. Las etiquetas para cada segmento de audio se encuentran con el nombre del archivo WAV correspondiente al siguiente patrón:

`AudioN_N-VXX-AYY-DZZ.wav`.

El nombre del archivo está dividido en cinco segmentos separados por subguion (_) y el símbolo de guión (-), donde,

- En `AudioN`, `N` indica el identificador numérico de la fuente de audio; este puede estar entre 1 y n.
- `N`, indica el número del segmento de audio que puede estar entre 1 y n.
- En `VXX`, La `V` se refiere a la dimensión emocional de valencia, y `XX` indica el valor numérico de la valencia de ese segmento, el cual puede estar entre 01 y 05.
- En `AYY`, La `A` se refiere a la dimensión emocional de excitación, y `YY` indica el valor numérico de la dimension emocional de excitación del segmento, el cual puede estar entre 01 y 05.
- En `DZZ`, La `D` se refiere a la dimensión emocional de dominancia, y `ZZ` indica el valor numérico de la dominancia de ese segmento, el cual puede estar entre 01 y 05.

Por ejemplo, el nombre del archivo `Audio1_1-02-02-04` indica que ese segmento pertenece al audio 1, el cual es el primer segmento de ese audio, y sus dimensiones emocionales son **Valencia** 02 (algo triste), **Excitación** 02 (algo calmado), y **Dominancia** 04 (algo dominante).

5.1.2. Validación técnica

Diversidad Emocional

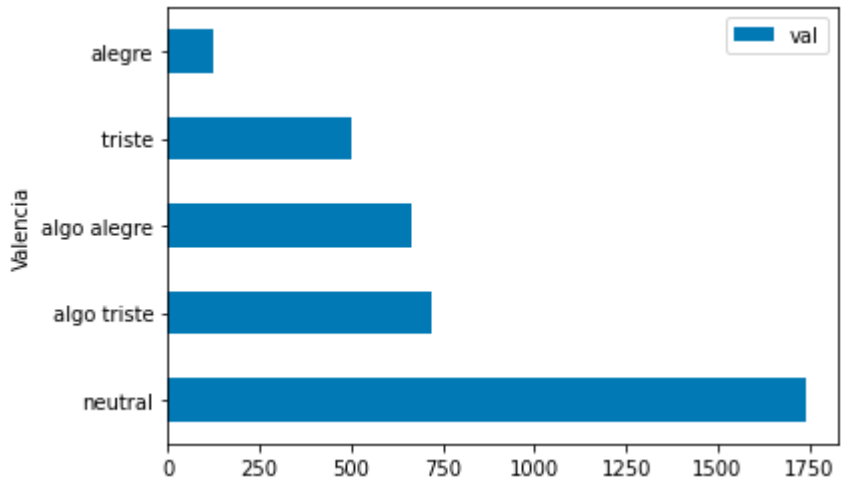
Para realizar la validación sobre la diversidad emocional del corpus, se siguieron los siguientes pasos:

1. Primero, una vez que se obtuvo el corpus, con las etiquetas globales, es decir, las etiquetas que se obtuvieron en la etapa del consenso, explicada, en el capítulo anterior, en la sección 4.5. Etiquetado, se organizaron las 11'247 etiquetas en los tres grupos de dimensiones emocionales, 3749 etiquetas para la dimensión emocional de la valencia, 3749 para la dimensión emocional de la excitación y 3749 para la dimensión emocional de la dominancia.
2. Segundo, se contaron, la cantidad de segmentos, por cada una de las escalas de las dimensiones emocionales, las cuáles se dividen en 5. Para la valencia de triste a alegre, excitación, de calma a activo y la dominancia, de dominado a dominante.

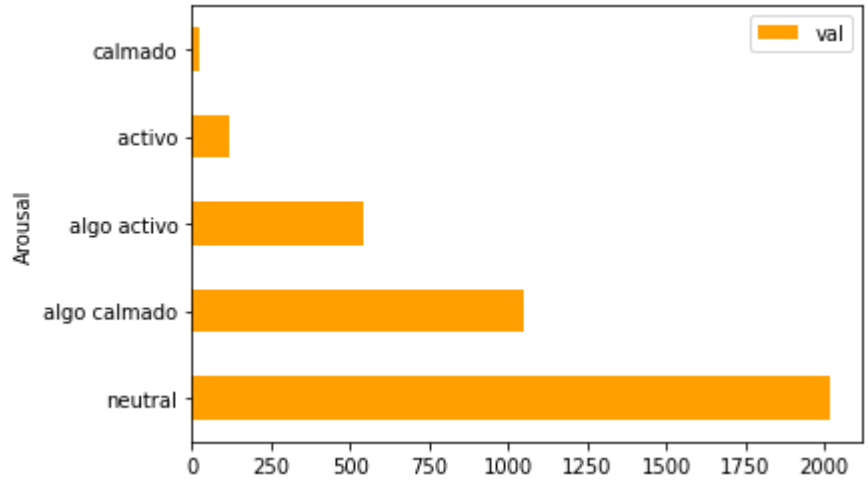
```
Tercero y último, se obtuvieron las siguientes cantidades por cada una de
    las escalas de las dimensiones emocionales, las cuáles se pueden
    observar en la Figura 5.1.
Para la valencia:
    Triste - 502 segmentos
    Algo triste - 721 segmentos
    Neutral - 1741 segmentos
    Algo alegre - 663 segmentos
    Alegre - 122 segmentos
Para la excitación:
    Calma - 21 segmentos
    Algo calmado - 1048 segmentos
    Neutral - 2020 segmentos
    Algo activo - 543 segmentos
    Activo - 117 segmentos
Para la dominancia:
    Dominado - 10 segmentos
    Algo dominado - 446 segmentos
    Neutral - 676 segmentos
    Algo dominante - 1660 segmentos
    Dominante - 957 segmentos
```

Listing 5.1: Pasos para la elaboración de la diversidad emocional del corpus

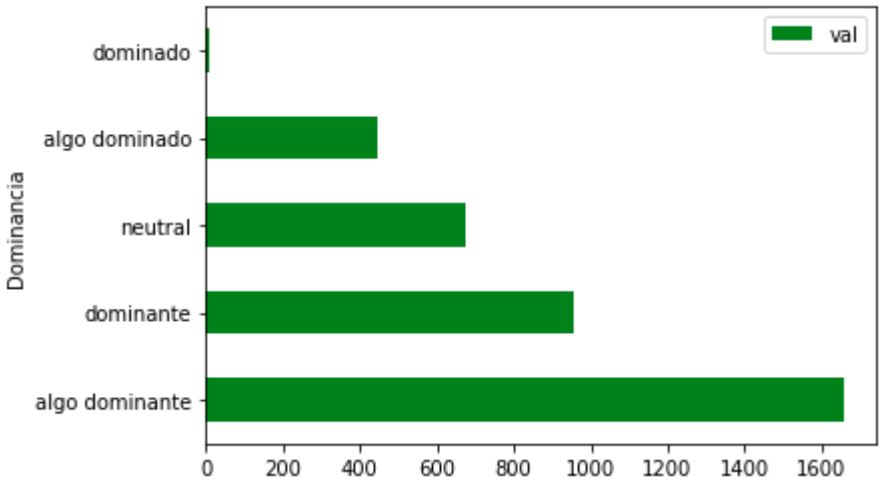
Así, se identificó, la diversidad emocional del contenido del corpus. La figura 5.1, muestra la distribución del contenido emocional del corpus, en términos de valencia, excitación y dominancia. Como se puede observar, no existe una distribución uniforme de las dimensiones emocionales, en las tres dimensiones. Tanto en la dimensión de valencia como excitación, predomina el estado neutral a diferencia, de la dominancia, ya que el estado que predomina, es algo dominante. En la dimensión de la valencia, se puede observar una distribución más homogénea entre los estados algo triste y algo alegre, y después, del estado neutral, se observa, que las dimensiones, algo triste y triste, son las que predominan. En el caso de la dimensión de excitación, después del estado neutral, predomina el estado algo calmado y finalmente en la dominancia, predominan los estados positivos, algo dominante y dominante.



(a) Valencia.



(b) Excitación.



(c) Dominancia.

Figura 5.1: Histograma de la distribución de las dimensiones emocionales en el corpus. Fuente propia.

5.1.3. Coincidencia de etiquetado

Se compara la concordancia entre el etiquetado individual de cada uno de los etiquetadores en el corpus utilizando el método Alfa de Cronbach. Se calculó la concordancia total en cada dimensión emocional.

A continuación, se puede observar, como se realizó el proceso de cálculo del método Alfa de Conbach, para las etiquetas, realizadas por los tres etiquetadores, de cada una de las dimensiones emocionales.

Donde, K, representa el número de segmentos que tiene el corpus, representado por 9 vectores, de 3749 etiquetas cada uno. Los tres primeros vectores, representan las etiquetas realizadas por los 3 etiquetadores, con la dimensión emocional, valencia, vector 1 (3749 etiquetas, etiquetador 1), vector 2 (3749 etiquetas, etiquetador 2) y vector 3 (3749 etiquetas, etiquetador 3). Los siguientes vectores, del 4 al 6, representan a la dimensión emocional de excitación, vector 4 (3749 etiquetas, etiquetador 1) vector 5 (3749 etiquetas, etiquetador 2) y vector 6 (3749 etiquetas, etiquetador 3). Finalmente, los vectores, del 7 al 9, hacen referencia, a la dimensión emocional, dominancia, vector 7 (3749 etiquetas, etiquetador 1), vector 8 (3749 etiquetas, etiquetador 2) y vector 9 (3749 etiquetas, etiquetador 3) . V_i hace referencia, a la varianza de cada uno de los ítems, de cada una de las dimensiones emocionales, por cada uno de los tres etiquetadores, y se calcula, con la fórmula de la varianza, en función a las etiquetas realizadas, por cada uno de los tres etiquetadores. Finalmente, V_t, hace referencia, a la varianza total, que representa, la varianza de la suma de las 3749 etiquetas, por cada una de las dimensiones emocionales, de los tres etiquetadores.

$AlphaConbach_V$ hace referencia, al valor obtenido del cálculo de la métrica alfa de cronbach, de la dimensión emocional de valencia:

$$AlphaCronbach_V = \frac{K}{(K-1)} \cdot \frac{(1 - V_i)}{V_t}$$

$$AlphaCronbach_V = \frac{3749}{(3749-1)} \cdot \frac{(1 - 850{,}66)}{78500{,}22}$$

$$AlphaCronbach_V = 0{,}9894$$

$AlphaCronbach_A$ hace referencia, al valor obtenido, del cálculo de la métrica alfa de cronbach, de la dimensión emocional, de excitación:

$$AlphaCronbach_A = \frac{K}{(K-1)} \cdot \frac{(1-V_i)}{V_t}$$

$$AlphaCronbach_A = \frac{3749}{(3749-1)} \cdot \frac{(1-3685{,}1)}{7494557{,}55}$$

$$AlphaCronbach_A = 0{,}9998$$

$AlphaCronbach_V$ hace referencia, al valor obtenido, del cálculo de la métrica alfa de cronbach, de la dimensión emocional, de dominancia:

$$AlphaCronbach_D = \frac{K}{(K-1)} \cdot \frac{(1-V_i)}{V_t}$$

$$AlphaCronbach_D = \frac{3749}{(3749-1)} \cdot \frac{(1-3430{,}94)}{4222272{,}89}$$

$$AlphaCronbach_D = 0{,}9995$$

El cuadro 5.3, presenta el valor alfa de cronbach, para cada dimensión emocional. En los tres casos, el valor, es superior a 0,9,lo que significa, que las etiquetas, entre los tres etiquetadores, tienen una concordancia interna alta.

Valencia	Excitación	Dominancia
0.9894	0.9998	0.9995

Cuadro 5.3: *Alfa Cronbach* por cada dimensión emocional.

5.1.4. Discusión

Tal cómo se menciona en la sección 2.7 del Marco Teórico, en la investigación realizada por [22] se definieron cuatro aspectos a ser considerados en el diseño de una base de datos oral de emociones, los cuáles son: alcance, naturalidad, contexto y descriptores. En esta sección se analizará el corpus construido en base a estos términos, para ello se han propuesto los siguientes criterios de evaluación para cada aspecto.

Categoría 1: Alcance

CRIT-1.1: El lenguaje español peruano presenta distintos dialectos según los orígenes de cada una de las personas, es por ello que los audios seleccionados para la construcción del corpus deben presentar al menos más de un sólo tipo de dialecto.

ID	Nombre	Tipo Dialecto
1	Debate Presidencial	Estándar
2	Testimonios de Flor Huilca y Martha Flores	Estándar
3	Testimonio de la Calificación al Mundial de Perú 2018	Estándar
4	Reportaje de tragedia en Arequipa 1996	Estándar
5	Reportaje de heladas en Arequipa	Vernacular
6	Reportaje de Huaynacotas en Arequipa	Vernacular
7	Entrevista de ataque de UFO a piloto de la FAP en La Joya, Arequipa	Standard
8	Testimonio de Maria Antonieta Quispe	Estándar
9	Reportaje de fin de semana en Arequipa	Estándar
10	Entrevista ”Derechos sexuales y reproductivos en Perú”	Estándar
11	Reportaje muerte Maryori en Huancayo	Vernácular
12	Charla TedxCharacato Mónica Huerta	Estándar
13	Reportaje de restaurante peruano en México	Estándar
14	Testimonio Cipriana Huamani	Estándar

Cuadro 5.4: CRIT 1.1: Tipos de Dialecto

Cómo se puede observar en el Cuadro 5.4 predomina el tipo de dialecto estándar, que

hace referencia a una variedad de la lengua difundida y entendida por la mayoría de los hablantes, utilizada frecuentemente en la educación formal y medios de comunicación. En el corpus diseñado representa el 79 %, sin embargo también se puede observar la presencia del tipo de dialecto vernacular, el cuál hace referencia al dialecto que es nativo, originario, autóctono del lugar, y que representa el 21 % en el corpus diseñado, lo cuál enriquece el corpus generado.

CRIT-1.2: La muestra de la población debe presentar un equilibrio entre el género masculino y femenino en los audios seleccionados, ya que según [28] se ha observado que el género puede afectar en la expresión de las emociones. Además según los estudios realizados en [42] encuentran que las mujeres son más expresivas que los hombres.

Se observa según el Cuadro 5.2 que el total de participantes es de 80, entre los cuales 44 son hombres y 36 mujeres, tal como se ve en la imagen 5.2. Si bien la distribución no es homogénea la diferencia entre ambos géneros es considerable.

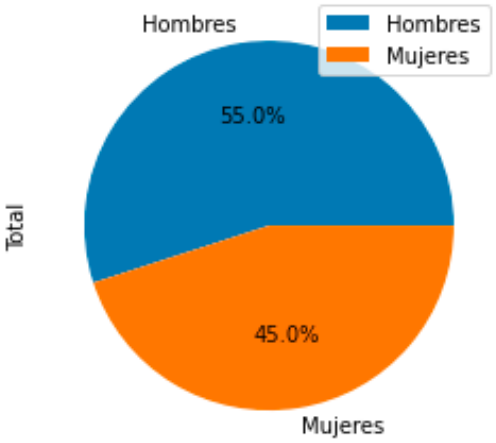


Figura 5.2: Distribución del género en el corpus. Fuente propia.

CRIT-1.3: Los corpus orales de emociones que tienen como objetivo abarcar el rango total de las emociones, tienen alrededor de 12 horas de material emocional. Si bien el corpus diseñado presenta 7 horas, 45 minutos y 52 segundos de audio, el cuál no cumple con el estándar indicado en [22], alcanza un 66 % del objetivo estándar, lo cuál es considerable para poder utilizarlo en modelos de *Deep Learning*.

CRIT-1.4: El número de participantes sugerido en [22] debe ser superior a diez.

En el diseño y construcción del corpus se utilizó un total de ochenta participantes, los cuáles están distribuidos en los audios/videos seleccionados.

CRIT-1.5: El propósito de la elaboración del corpus es el reconocimiento de las emociones utilizando el idioma español con acento peruano, de modo que se pueda proporcionar un punto de partida adecuado para entrenar clasificadores robustos y modelos emocionales. Dicho criterio es validado en la sección de evaluaciones, específicamente la evaluación cuantitativa.

Categoría 2: Naturalidad

CRIT-2.1: ¿Cómo se intentó mantener el equilibrio de las emociones en el diseño del corpus? En el diseño y construcción del corpus, se intentó equilibrar la presencia de emociones, tanto positivas como negativas, realizando la búsqueda de audios, mediante palabras claves, relacionadas a sucesos ocurridos en nuestro país, que fueran alegres, como nuestra cultura y folclore, o tristes, como testimonios del terrorismo, pobreza y el crimen o también polémicos, como sucesos extraños, debates, esterilización forzada. A pesar de ello, como se menciona en [22, 16, 49, 58] el precio de la naturalidad, es la falta de control. La selección de audios, con emociones representativas, no permitió obtener el resultado esperado, tal como se aprecia en la figura 5.1, en la que se observa, que predomina, el estado emocional neutral, tanto en las dimensiones emocionales, de valencia y excitación. Sin embargo, en la dominancia, predomina la característica algo dominante. Según lo indicado en [22, 17, 49], una respuesta útil, a la complejidad natural, es retirarse del ideal de cubrir todo el dominio de las emociones y centrarse en una subregión específica. Cómo se observa en la dimensión de la valencia, donde predomina, después del estado neutral, el estado emocional negativo (algo triste y triste), en la excitación, el estado emocional, algo calmado y en la dominancia los estados emocionales positivos (algo dominante y dominante).

En el corpus, se hizo uso de audios, cuidadosamente seleccionados, el cuál permitió observar la representación genuina de las emociones, excluyendo, monólogos o discursos leídos.

CRIT-2.2: ¿Cómo sería etiquetado el corpus diseñado? Según [61], este corpus, sería etiquetado como natural, ya que los participantes, se expresan libremente. La ventaja de un corpus oral de emociones de este tipo, es que permite su uso en entornos reales,

ya que esta basado en situaciones genuinas.

Categoría 3: Contexto

CRIT-3.1: Los escenarios, de los corpus emocionales, revisados en [22, 16, 58, 61], están basados, en expresiones aisladas o diálogos breves, estos escenarios, eliminan el contexto, el cual, es un componente importante. En este corpus, la duración media de los diálogos es de aproximadamente treinta minutos, lo que permite contextualizar las señales y el fluir de las emociones. Dado, que el material, fue seleccionado adecuadamente desde una perspectiva basada en la obtención de las emociones mediante diálogos, entrevistas y conversaciones genuinas, se puede analizar las emociones en un contexto adecuado. Una vez finalizado el proceso de diseño y construcción del corpus, se realizaron las evaluaciones emocionales, de modo, que los etiquetadores puedan juzgar el contenido emocional, en base al desarrollo secuencial de los diálogos.

Categoría 4: Descriptores

CRIT-4.1: ¿Las categorías emocionales consideradas representan el contenido emocional del corpus? El estudio realizado en [22], nos indica que para los corpus actuados, las emociones categóricas, son la mejor opción, ya que, se sabe de antemano cuál será la emoción que debe tener cada oración del diálogo, sin embargo, para los corpus naturales, la asignación de emociones, es más compleja, ya que involucra picos emocionales, por lo que, están mejor representados a través de la continuidad de las emociones, en vez de utilizar una emoción específica, ya que se puede capturar aspectos complementarios de la manifestación emocional. Por ello, se vio por conveniente, utilizar las dimensiones emocionales de valencia, excitación y dominancia, ya que representan adecuadamente una aproximación razonable al contenido de las emociones observadas en el corpus, mejorando, la descripción emocional del corpus recopilado.

5.2. Evaluación cuantitativa

5.2.1. Eficacia real del corpus en clasificadores de *Deep Learning*

La evaluación del corpus en modelos de *Deep Learning* permitirá validar la eficacia real del mismo, analizando el grado en que el corpus diseñado alcanza su objetivo de reconocimiento de emociones.

Para ello se utilizó específicamente una red neuronal convolucional y una red neuronal residual (ResNet), adecuadas para el procesamiento de imágenes, considerando que cada uno de los segmentos de audio es convertido a un espectrograma para identificar las características MFCCs.

5.2.2. Resultados de la evaluación

La selección de las funciones de coste MSE y CCC se basaron en el análisis de la literatura [10] debido a que el reconocimiento de dimensiones emocionales es una tarea más compleja que solo el reconocimiento de emociones categóricas, ya que se convierte en aprendizaje multitarea debido a que el clasificador no solo debe predecir un valor sino tres valores al mismo tiempo.

Los cuadros 5.5 y 5.6 muestran los resultados utilizando las dos funciones de pérdida en el mismo conjunto de datos con las dos arquitecturas de red.

El cuadro 5.5 muestra los resultados obtenidos al realizar el entrenamiento de las redes neuronales: convolucional y Resnet, considerando que el rango de la función CCC oscila entre -1 perfecto desacuerdo y 1 perfecto acuerdo entre la predicción del modelo y la etiqueta real, utilizando la función CCC loss el puntaje obtenido para las tres dimensiones emocionales es el más alto comparado con la función MSE, así mismo la media resultante con los resultados de la función CCC para ambos modelos de red neuronal es el mayor.

Del mismo modo en el cuadro 5.6 se pueden observar que los resultados obtenidos con la función CCC para ambos modelos de red neuronal es el mayor, tanto para las dimensiones emocionales como para la media resultante.

Red Neuronal	Loss	V	A	D	Media
CNNDim-3462	CCCL	0.838	0.837	0.848	0.842
CNNDim-3497	MSE	0.774	0.623	0.748	0.688
Resnet-2163	CCCL	0.794	0.737	0.817	0.775
Resnet-2179	MSE	0.763	0.633	0.768	0.700

Cuadro 5.5: Resultados de la evaluación del entrenamiento con las funciones de pérdida en el corpus.

Red Neuronal	Loss	V	A	D	Media
CNNDim-3462	CCCL	0.843	0.644	0.799	0.726
CNNDim-3497	MSE	0.826	0.615	0.799	0.710
Resnet-2163	CCCL	0.757	0.566	0.737	0.654
Resnet-2179	MSE	0.705	0.544	0.712	0.627

Cuadro 5.6: Resultados de la evaluación de la validación con las funciones de pérdida en el corpus.

En la partición de entrenamiento la media de las dimensiones emocionales fue de **0.842** y en la partición de validación la media de las dimensiones emocionales fue de **0.726**, estos resultados sugieren que los modelos empleados en la evaluación han tenido una predicción significativamente buena con el corpus diseñado en esta tesis.

5.2.3. Discusión

El principal objetivo de la realización de la evaluación cuantitativa es poder medir la eficacia real del corpus, es decir, el grado en que el corpus alcanza sus objetivos, específicamente el reconocimiento de emociones utilizando el idioma español con acento peruano. Para ello se realizó la evaluación del corpus en los modelos de deep learning CNN y ResNet, utilizando la función de pérdida CCC, la cuál fue seleccionada de entre las demás funciones de pérdida MSE, MAE, ya que según la literatura obtuvo mejores resultados en el reconocimiento de emociones dimensionales. Lo que en esta tesis también queda demostrado ya que se obtuvieron resultados de predicción significativamente buenos en cada una de las dimensiones emocionales.

Analizando los resultados se puede observar que la dimensión emocional que obtiene el mayor valor de predicción es la dominancia y esto se debe a la relación directa que existe entre la mayor cantidad de muestras que posee esta dimensión en sus diferentes escalas sobretodo la escala algo dominante, a diferencia de la valencia y excitación en las que predomina la escala neutral. A pesar de ello dichas dimensiones obtienen un resultado por encima de 0.5.

Los resultados obtenidos en la evaluación cuantitativa tienen una relación directa con la evaluación cualitativa del corpus, ya que en ella se analiza cada uno de los

aspectos involucrados en su diseño y construcción, el cuál después se puede medir en la evaluación cuantitativa, específicamente en uno de los criterios analizados en la evaluación cualitativa que fue el alcance, ya que se analiza el propósito del corpus, el cuál, como ya se ha mencionado, es el reconocimiento de las emociones en el lenguaje español con acento peruano, dicho reconocimiento se midió empleando los modelos de deep learning y obteniendo el grado de predicción para cada dimensión emocional.

5.3. Limitaciones

Entre las limitaciones del corpus, resaltan las siguientes: La primera es el tamaño, que hace referencia a la duración total del corpus en cuánto al tiempo y al número de segmentos, ya que según el estudio realizado en el estado del arte, existen corpus como *Design and Evaluation of Adult Emotional Speech Corpus for Natural Environment* [58] que cuenta con 13'500 segmentos y *the MSP-PODCAST CORPUS* [48] que tiene una duración de 27 horas, a comparación de estos corpus el corpus diseñado y construido es considerablemente más pequeño en lo que respecta al tiempo y número de segmentos. Este aspecto podría afectar su rendimiento en sistemas de reconocimiento de emociones basados en modelos de Deep Learning, los cuáles requieren grandes cantidades de datos para las etapas de entrenamiento y validación. Sin embargo esta limitación podría aliviarse aplicando operaciones de aumento de datos.

La segunda limitación es la distribución heterogénea de las dimensiones emocionales. Aunque durante la etapa de diseño del corpus se buscó equilibrarlo emocionalmente, el resultado no fue tal, inclusive se puede observar que el estado emocional predominante para la valencia y excitación es el neutral, lo que podría ocasionar la falta de precisión en los sistemas de reconocimiento de emociones para nuevas muestras que se encuentren en los vacíos del espacio emocional del corpus. A pesar de ello según el análisis realizado en [34] la distribución emocional obtenida, es similar a los resultados observados en otros corpus orales de emociones naturales.

La tercera y última limitación fue el uso de sólo tres etiquetadores, lo cual permitió obtener un valor considerable en la evaluación *Alfa de Cronbach*, sin embargo no implica que estemos libres de sesgos o que todos los segmentos hayan sido bien etiquetados.

Capítulo 6

Conclusiones y trabajo futuro

6.1. Conclusiones

El corpus presentado en esta tesis está conformado por 3749 segmentos de audio, haciendo un total de 7 horas 45 minutos y 52 segundos de duración. Tiene voces de 80 participantes interactuando en escenarios naturales tales como debates, entrevistas y reportajes. Todos los participantes son adultos que tienen el español con acento peruano como lengua materna. El etiquetado de los tres atributos emocionales (valencia, excitación y dominancia) fue realizado por tres anotadores y se obtuvo una concordancia interna de 0.9 para el coeficiente de Alfa de Cronbach.

El corpus fue evaluado mediante un método cualitativo que permitió medir el alcance, naturalidad, contexto y sus descriptores. En cuanto al alcance se analizaron los criterios del tipo de dialecto, en el que predominó el dialecto estándar versus el vernacular, el equilibrio entre géneros, donde predominó el género masculino con 44 participantes versus 36 participantes femeninos, el número de horas, el número de participantes y el propósito del corpus. Con respecto a la naturalidad se analizó la forma en que se intento mantener la diversidad emocional, la cuál no fue homogénea. Respecto al contexto se analizaron los escenarios utilizados en los audios seleccionados, los cuáles estuvieron basados en situaciones reales con expresión genuina de las emociones. Finalmente, los descriptores utilizados en el corpus fueron las dimensiones emocionales

de valencia, excitación y dominancia, ya que según la literatura representan mejor a los corpus naturales.

También, se realizó una evaluación cuantitativa del corpus para medir su eficacia real. Fueron utilizadas una red neuronal convolucional y una red neuronal residual (ResNet). El desempeño de los clasificadores para la etapa de entrenamiento fue de 0.842 y 0.775 respectivamente y para la etapa de validación se obtuvieron los valores 0.726 y 0.654, respectivamente. Considerando la métrica CCC utilizada, donde el valor de 1 representa la predicción perfecta, los valores obtenidos fueron significativos en cuanto a la predicción realizada con ambos clasificadores.

6.2. Trabajo futuro

Como trabajo futuro se propone incrementar el tamaño del corpus debido a que cuenta con una cantidad de horas (7 horas 45 minutos) la cuál no llega a sobrepasar el tamaño de los corpus emocionales más populares, por ejemplo, el MSP-IMPROV (9 horas), el IEMOCAP (12 horas) y el MSP-Podcast Corpus (27 horas).

Otra línea de trabajo interesante es la automatización total de la etapa de segmentación. Esta automatización debería permitir recuperar segmentos del habla cumpliendo algunos requerimientos, por ejemplo, tener solo un hablante por segmento, que el segmento tenga una duración mínima de 3 segundos y una duración máxima de 15 segundos y finalmente que el segmento exprese una sola emoción, ya sea una palabra o frase. De este modo, el proceso podría ser replicado en audios obtenidos en tiempo real y procesados sin intervención humana.

Bibliografía

[1] et al Abhinav Jain. Kdd 2020 tutorial: Overview and importance of data quality for machine learning tasks. *Overview and Importance of Data Quality for Machine Learning Tasks*, 59, 01 2020.

[2] S. Abrilian, Laurence Devillers, Stéphanie Buisine, and Jean-Claude Martin. Emotv1: Annotation of real-life emotions for the specifications of multimodal a ective interfaces. 01 2005.

[3] Shazia Afzal, Rajmohan C, Manish Kesarwani, Sameep Mehta, and Hima Patel. Data readiness report. *CoRR*, abs/2010.07213, 2020.

[4] Murray Alpert, Enrique Pouget, and Raul Silva. Reflections of depression in acoustic measures of the patient's speech. *Journal of affective disorders*, 66:59–69, 10 2001.

[5] Noam Amir. Analysis of an emotional speech corpus in hebrew based on objective criteria this volume. 2000.

[6] Noam Amir, Samuel Ron, and Nathaniel Laor. Analysis of an emotional speech corpus in hebrew based on objective criteria. *Proc. ISCA Workshop (ITRW) Speech and Emotion*, 01 2000.

[7] Jeremy Ang, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg, and Andreas Stolcke. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. *ICSLP 2002*, 10 2002.

[8] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *CoRR*, abs/1912.06670, 2019.

[9] Magda B. Arnold. *Emotion and Personality Volume II : Neurological and Physiological Aspects*, page 430. 06 1963.

[10] Bagus Tris Atmaja and Masato Akagi. Evaluation of error and correlation-based loss functions for multitask learning dimensional speech emotion recognition, 2020.

[11] Moataz Ayadi, Mohamed S. Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44:572–587, 03 2011.

[12] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[13] Anton Batliner, Björn Schuller, Dino Seppi, Stefan Steidl, Laurence Devillers, Laurence Vidrascu, Thurid Vogt, Vered Aharonson, and Noam Amir. *The Automatic Recognition of Emotions in Speech*, pages 71–99. 01 2011.

[14] Alan W Black. Cmu wilderness multilingual speech dataset. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975, 2019.

[15] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, December 2008.

[16] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[17] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed Abdelwahab, Najmeh Sadoughi, and Emily Mower Provost. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8:1–1, 01 2016.

[18] J. Chen, Chenhui Wang, Ke jun Wang, Chaoqun Yin, Cong Zhao, Tao Xu, Xinyi Zhang, Ziqiang Huang, Meichen Liu, and Tao Yang. Heu emotion: A large-scale database for multi-modal emotion recognition in the wild. *Neural Comput. Appl.*, 33:8669–8685, 2021.

[19] Diane Cook, Glen Duncan, Gina Sprint, and Roschelle Fritz. Using smart city technology to make healthcare smarter. *Proceedings of the IEEE*, PP:1–15, 01 2018.

[20] Roddy Cowie and E. Douglas-Cowie. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. pages 1989 – 1992 vol.3, 11 1996.

[21] Escudero Mancebo David, Aguilar Lourde, Gonzalez Ferreras Cesar, Vivaracho Pascual, and Cardenoso Payo Valentin. Caracterizacion acustica del acento basada en corpus: un enfoque multilingue ingles-espanol. In *International Conference on Machine Learning*, page 10. Departamento de Informatica, Universidad de Valladolid, Espana Departamento de Filología Hispanica, Universidad Autonoma de Barcelona, Espana, 10 2011.

[22] Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie, and Peter Roach. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40:33–60, 04 2003.

[23] Paul Ekman, Wallace V. Friesen, and Phoebe Ellsworth. Introduction. In Paul Ekman, Wallace V. Friesen, and Phoebe Ellsworth, editors, *Emotion in the Human Face*, volume 11 of *Pergamon General Psychology Series*, pages 1–6. Pergamon, 1972.

[24] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587, 2011.

[25] Humberto Espinosa, Juan Martínez-Miranda, Ismael Espinosa Curiel, Josefina Rodríguez-Jacobo, Luis Villaseñor-Pineda, and Himer Avila-George. Iesc-child: An interactive emotional children's speech corpus. *Computer Speech and Language*, 59, 01 2020.

[26] Weiquan Fan, Xiangmin Xu, Xiaofen Xing, Weidong Chen, and Dongyan Huang. LSSED: a large-scale dataset and benchmark for speech emotion recognition. *CoRR*, abs/2102.01754, 2021.

[27] Raul Fernandez and Rosalind W. Picard. Modeling drivers' speech under stress. *Speech Communication*, 40(1):145–159, 2003.

[28] Agneta Fischer, P. Rodriguez Mosquera, Annelies van Vianen, and Antony Manstead. Gender and culture differences in emotion. *Emotion (Washington, D.C.)*, 4:87–94, 04 2004.

[29] Leidy Paola Bolaños Florido. El estudio socio-histórico de las emociones y los sentimientos en las ciencias sociales del siglo xx. *Revista de Estudios Sociales*, 55:178–191, 01 2016.

[30] D.J. France, R.G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, 47(7):829–837, 2000.

[31] Nico H. Frijda. *The Emotions*. Cambridge University Press, 1986.

[32] G.M. Gonzalez and Julian Samora Research Institute. *Bilingual Computer-assisted Psychological Assessment: An Innovative Approach for Screening Depression in Chicanos/Latinos*. Latino studies series. Julian Samora Research Institute, Michigan State University, 1999.

[33] Jeffrey A. Gray. Introduction. In Jeffrey A. Gray, editor, *The Neuropsychology of Anxiety*, volume 69 of *Brit. J. Psychol*, pages 417–434. Brit. J. Psychol, 1982.

[34] Michael Grimm, Kristian Kroschel, Emily Mower, and Shrikanth Narayanan. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10):787–800, 2007. Intrinsic Speech Variations.

[35] John HL Hansen and Gang Liu. Unsupervised accent classification for deep data fusion of accent and language information. *Speech Communication*, 78:19–33, 2016.

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[37] Akemi Iida, Nick Campbell, Soichiro Iga, Fumito Higuchi, and Michiaki Yasumura. A speech synthesis system with emotion for assisting communication. 01 2000.

[38] Carroll E. Izard. *The face of emotion*, page 430. 06 1971.

[39] Arnaldo Cândido Júnior, Edresson Casanova, Anderson da Silva Soares, Frederico Santos de Oliveira, Lucas Oliveira, Ricardo Corso Fernandes Junior, Daniel Peixoto Pinto da Silva, Fernando Gorgulho Fayet, Bruno Baldissera Carlotto, Lucas Rafael Stefanel Gris, and Sandra Maria Aluísio. CORAA: a large corpus of spontaneous and prepared speech manually validated for speech recognition in brazilian portuguese. *CoRR*, abs/2110.15731, 2021.

[40] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345, 2019.

[41] Eun Ho Kim, Kyung Hak Hyun, Soo Hyun Kim, and Yoon Keun Kwak. Speech emotion recognition using eigen-fft in clean and noisy environments. In *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*, pages 689–694, 2007.

[42] Ann M. Kring and Austin Gordon. Sex differences in emotion: expression, experience, and physiology. *Journal of personality and social psychology*, 74 3:686–703, 1998.

[43] Muddasar Laghari, Muhammad Junaid Tahir, Abdullah Azeem, Waqar Riaz, and Yi Zhou. Robust speech emotion recognition for sindhi language based on deep convolutional neural network. In *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*, pages 543–548. IEEE, 2021.

[44] Siddique Latif, Adnan Qayyum, Muhammad Usman, and Junaid Qadir. Cross lingual speech emotion recognition: Urdu vs. western languages. In *2018 International Conference on Frontiers of Information Technology (FIT)*, pages 88–93. IEEE, 2018.

[45] Chul Min Lee and S.S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303, 2005.

[46] Rensis Likert. *A technique for the measurement of attitudes / by Rensis Likert.* Archives of psychology ; no. 140. [s.n.], New York, 1985 - 1932.

[47] Llisterri, Joaquim. Los elementos suprasegmentales., 2016.

[48] Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, 2019.

[49] Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, 2019.

[50] Elena Lyakso, Olga Frolova, Arman Kaliyev, Viktor Gorodnyi, Aleksey Grigorev, and Yuri Matveev. Ad-child.ru: Speech corpus for russian children with atypical development. In Albert Ali Salah, Alexey Karpov, and Rodmonga Potapova, editors, *Speech and Computer*, pages 299–308, Cham, 2019. Springer International Publishing.

[51] Veronika Makarova and Valery Petrushin. Ruslana: a database of russian emotional utterances. volume 1, 01 2002.

[52] William McDougall. *An Introduction to Social Psychology*, page 559. 06 2017.

[53] Andrew McStay. *Emotional AI: The rise of empathic media*. Sage, 2018.

[54] Ali Hamid Meftah, Yousef Ajami Alotaibi, and Sid-Ahmed Selouani. Evaluation of an arabic speech corpus of emotions: A perceptual and statistical analysis. *IEEE Access*, 6:72845–72861, 2018.

[55] Alexis Mendoza, Alvaro Cuno, Nelly Condori-Fernandez, and Wilber Ramos Lovón. An evaluation of physiological public datasets for emotion recognition systems. In Juan Antonio Lossio-Ventura, Jorge Carlos Valverde-Rebaza, Eduardo Díaz, and Hugo Alatrista-Salas, editors, *Information Management and Big Data*, pages 90–104, Cham, 2019. Springer International Publishing.

[56] Juan M. Montero, Juana M. Gutiérrez-Arriola, and José Colás. Analysis and modelling of emotional speech in spanish. 1999.

[57] NIH. What is voice? what is speech? what is language? `https://www.nidcd.nih.gov/health/what-is-voice-speech-language`, Dec 2020.

[58] Jia Ning, Chunjun Zheng, and Wei Sun. Design and evaluation of adult emotional speech corpus for natural environment. pages 53–56, 08 2020.

[59] Varun Boyanapally Nitesh Varma Rudraraju. Data quality model for machine learning (dissertation). In *Data Quality Model for Machine Learning*, page 107. Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden, 05 2019.

[60] Keith Oatley and P. Johnson-laird. Towards a cognitive theory of emotions. *Cognition and Emotion*, 1:29–50, 03 1987.

[61] Sheetal Patil and G. K. Kharate. A review on emotional speech recognition: Resources, features, and classifiers. In *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, pages 669–674, 2020.

[62] Sheetal Patil and Gajanan K. Kharate. A review on emotional speech recognition: Resources, features, and classifiers. *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, pages 669–674, 2020.

[63] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[64] Humberto Pérez-Espinosa, Juan Martínez-Miranda, Ismael Espinosa-Curiel, Josefina Rodríguez-Jacobo, Luis Villaseñor-Pineda, and Himer Avila-George. Iescchild: An interactive emotional children's speech corpus. *Computer Speech & Language*, 59:55–74, 2020.

[65] Valery Petrushin. Emotion in speech: Recognition and application to call centers. *Proceedings of Artificial Neural Networks in Engineering*, 01 2000.

[66] Sylvaine Picard, Camille Chapdelaine, Cyril Cappi, Laurent Gardes, Eric Jenn, Baptiste Lefèvre, and Thomas Soumarmon. Ensuring dataset quality for machine learning certification, 11 2020.

[67] Robert Plutchik. Chapter 1 - a general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press, 1980.

[68] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. MLS: A large-scale multilingual dataset for speech research. In *Interspeech 2020*. ISCA, oct 2020.

[69] Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. The multilingual tedx corpus for speech recognition and translation, 2021.

[70] Ignasi Sanz, Roger Guaus, Angel Rodrguez, Patrícia Lázaro Pernias, Norminanda Vilar, Josep Maria Pont, Dolors Bernadas, Josep Oliver, Daniel Tena, and Ludovico Longhi. Validation of an acoustical modelling of emotional expression in spanish using speech synthesis techniques. 03 2001.

[71] Klaus Scherer. Psychological models of emotion. *The Neuropsychology of Emotion*, 01 2000.

[72] Klaus Scherer, Didier Grandjean, Tom Johnstone, Gudrun Klasmeyer, and Tanja Bänziger. Acoustic correlates of task load and stress. 01 2002.

[73] Florian Schiel, Silke Steininger, and Ulrich Türk. The SmartKom multimodal corpus at BAS. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain, May 2002. European Language Resources Association (ELRA).

[74] Björn Schuller and Anton Batliner. Computational paralinguistics: Emotion, affect and personality in speech and language processing. 2013.

[75] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Commun.*, 53:1062–1087, 2011.

[76] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39, 2013. Special issue on Paralinguistics in Naturalistic Speech and Language.

[77] Malcolm Slaney and Gerald McRoberts. Babyears: A recognition system for affective vocalizations. *Speech Communication*, 39(3):367–384, 2003.

[78] ZHAO Li SONG Peng, ZHENG Wenming. Joint subspace learning and feature selection method for speech emotion recognition. *Journal of Tsinghua University(Science and Technology)*, 58(4):347, 2018.

[79] Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. JSUT corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. *CoRR*, abs/1711.00354, 2017.

[80] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[81] Richard Stibbard. Automated extraction of tobi annotation data from the reading/leeds emotional speech corpus. *Proceedings of the ISCA ITRW on Speech and Emotion*, 09 2000.

[82] Franci Suni Lopez and Nelly Condori-Fernández. Design of an adaptive persuasive mobile application for stimulating the medication adherence. volume 178, 11 2017.

[83] Silvan S Tomkins. Affect theory. *Approaches to emotion*, pages 163–195, 01 1984.

[84] Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181, 2006.

[85] Ravichander Vipperla, Steve Renals, and Joe Frankel. Ageing voices: The effect of changes in voice parameters on asr performance. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010.

[86] Claude Vloeberghs, Patrick Verlinde, Carl Swail, Herman Steeneken, and Allan South. The impact of speech under ”stress.ºn military speech technology. (l’impact de la parole en condition de ”stress”sur less technologies vocales militaries). page 112, 03 2000.

[87] Changhan Wang, Juan Miguel Pino, Anne Wu, and Jiatao Gu. Covost: A diverse multilingual speech-to-text translation corpus. *CoRR*, abs/2002.01320, 2020.

[88] Zijie J. Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Polo Chau. CNN explainer: Learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1396–1406, feb 2021.

[89] Bernard Weiner and Sandra Graham. An attributional approach to emotional development. *Emotions, Cognition, and Behavior*, 3(4):167–191, 1984.

[90] Carl E. Williams and Kenneth N. Stevens. Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, 52(4B):1238–1250, 1972.

[91] Ben Williamson, Sian Bayne, and Suellen Shay. The datafication of teaching in higher education: critical issues and perspectives. *Teaching in Higher Education*, 25(4):351–365, 2020.

[92] Xuesong Yang, Kartik Audhkhasi, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, and Mark Hasegawa-Johnson. Joint modeling of accents and acoustics for multi-accent speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2018.

[93] Jinsung Yoon, Sercan Arik, and Tomas Pfister. Data valuation using reinforcement learning. In *International Conference on Machine Learning*, pages 10842–10851. PMLR, 2020.

[94] Nimra Zaheer, Obaid Ullah Ahmad, Ammar Ahmed, Muhammad Shehryar Khan, and Mudassir Shabbir. Semour: A scripted emotional speech repository for urdu. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2021.

[95] Marcely Zanon Boito, William Havard, Mahault Garnerin, Éric Le Ferrand, and Laurent Besacier. MaSS: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the Bible. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6486–6493, Marseille, France, May 2020. European Language Resources Association.

[96] Wisha Zehra, Abdul Rehman Javed, Zunera Jalil, Habib Ullah Khan, and Thippa Reddy Gadekallu. Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex & Intelligent Systems*, pages 1–10, 2021.

[97] Sebastian Zepf, Javier Hernandez, Alexander Schmitt, Wolfgang Minker, and Rosalind W. Picard. Driver emotion recognition for intelligent vehicles: A survey. *ACM Comput. Surv.*, 53(3), jun 2020.