**UNIVERSIDAD NACIONAL DE INGENIERIA**

**FACULTAD DE INGENIERÍA INDUSTRIAL Y DE SISTEMAS**



**TESIS**

**“DISEÑO DE APLICACIÓN DEL MODELO DE REDES NEURONALES (MRN) PARA RECONOCER EL ESTADO EMOCIONAL DE UNA VOZ GRABADA EN UN CALL CENTER”**

**PARA OBTENER EL TÍTULO PROFESIONAL DE**

**INGENIERA DE SISTEMAS**

**ELABORADO POR**

**TANIA MARIANELLA SOTOMAYOR FERNÁNDEZ**

**ASESOR: Ing. Rony Hancco**

**LIMA - PERÚ**

**2018**

*A Dios, a mis padres; hermanos y amigos.*

# ÍNDICE

[ÍNDICE 3](#_Toc531529999)

[LISTA DE FIGURAS 4](#_Toc531530000)

[LISTA DE TABLAS 5](#_Toc531530001)

[RESUMEN 6](#_Toc531530002)

[ABSTRACT 7](#_Toc531530003)

[CAPÍTULO I 8](#_Toc531530004)

[1.1. Generalidades 8](#_Toc531530005)

[1.2. Problemática 10](#_Toc531530006)

[**1.2.1. Formulación del problema** 12](#_Toc531530007)

[1.3. Objetivos 12](#_Toc531530008)

[1.3.1. Objetivo General 12](#_Toc531530009)

[1.3.2. Objetivos Específicos 13](#_Toc531530010)

[1.4. Hipótesis 13](#_Toc531530011)

[1.4.1. Hipótesis General: 13](#_Toc531530012)

[1.4.2. Hipótesis Específicas: 13](#_Toc531530013)

[1.5. Justificación 14](#_Toc531530014)

[CAPÍTULO II 15](#_Toc531530015)

[2.1. Antecedentes 15](#_Toc531530016)

[**2.1.1. “Multistage data selection based unsupervised speaker adaptation for personalized speech emotion recognition”** 15](#_Toc531530017)

[**2.1.2. “Emotion Recognition In Speech Signal”** 18](#_Toc531530018)

[**2.1.3. Detection and Analysis of Human Emotions through Voice and Speech Pattern Processing** 19](#_Toc531530019)

[2.2. Marco Teórico 21](#_Toc531530020)

[CAPÍTULO III 29](#_Toc531530021)

[CAPÍTULO IV 40](#_Toc531530022)

[ANEXOS 43](#_Toc531530023)

[Resultados del audio extraído por una persona fuera de la muestra 43](#_Toc531530024)

[BIBLIOGRAFÍA 45](#_Toc531530025)

# LISTA DE FIGURAS

[**Figura 1: Modelo Acústico Basado en el Ser** 17](#_Toc531530026)

[**Figura 2: Teoría de Excitación y Valencia** 23](#_Toc531530027)

[**Figura 3: Imagen de código en Python** 31](#_Toc531530028)

[**Figura 4: Emoción Triste** 32](#_Toc531530029)

[***Figura 5: Emoción Alegre*** 33](#_Toc531530030)

[**Figura 6: Emoción Enojado** 33](#_Toc531530031)

[**Figura 7: Emoción Neutro** 34](#_Toc531530032)

[**Figura 8: Rango de tiempo para las cuatro emociones** 35](#_Toc531530033)

[**Figura 9: Pitch del Estado Neutro** 36](#_Toc531530034)

[**Figura 10: Pitch del Estado Enojo** 36](#_Toc531530035)

[**Figura 11: Amplitud de audio de voz** 43](#_Toc531530036)

[**Figura 12: Emoción enojo de muestra extraída** 44](#_Toc531530037)

# LISTA DE TABLAS

[**Tabla 1: Información de GdC de Banco Falabella, San Isidro** 12](#_Toc531530039)

[**Tabla 2: Características del Estado Neutro** 24](#_Toc531530040)

[**Tabla 3: Características del Estado Enojado** 24](#_Toc531530041)

[**Tabla 4: Categoría de Emociones** 26](#_Toc531530042)

[***Tabla 5: Base de Datos de Emociones*** 30](#_Toc531530043)

[**Tabla 6: Matriz de confusión del Grupo Control** 39](#_Toc531530044)

[**Tabla 7: Matriz de confusión del Grupo Experimental** 39](#_Toc531530045)

[**Tabla 8: Resultados de Efectividad del Grupo Control** 40](#_Toc531530046)

[**Tabla 9: Resultados de Efectividad del Grupo Experimental** 40](#_Toc531530047)

[**Tabla 10: Resultados de aciertos y errores del Grupo Control** 41](#_Toc531530048)

[**Tabla 11: Resultados de aciertos y errores del Grupo Experimental** 41](#_Toc531530049)

# RESUMEN

Este trabajo propone la aplicación de las redes neuronales para reconocer el estado emocional de las clientes que buscan cancelar su tarjeta de crédito llamando a un call center de una entidad bancaria en el Perú. Dicho modelo ha sido estudiado y probado en anteriores investigaciones.

Como ya se hizo mención, en los antecedentes se muestran todos los algoritmos que se han usado en investigaciones pasadas y se propone un método basado en estos papers, pero dándole un enfoque más simplificado.

Además, se expone el problema que ha sido el origen de esta tesis, los objetivos que se esperan lograr, las hipótesis planteadas, así como los antecedentes que han precedido a este trabajo, un marco teórico para las definiciones que se mencionan en la presente investigación y finalmente se detalla el modelo de solución propuesto.

# ABSTRACT

This work proposes an application of neural networks to recognize the emotional state of customers who are looking for their credit card by calling a call center of a bank in Peru. This model has been studied and tested in previous investigations.

As already mentioned, in the background all the algorithms that have been used in the investigations are shown and a method based on these papers is proposed, but giving it a more simplified approach.

In addition, it explains the problem that has been the origin of this thesis, the objectives that are expected to achieve, the hypotheses raised, as well as the background that preceded this work, a theoretical framework to be mentioned in the present investigation. and finally, the proposed solution model is detailed.

# CAPÍTULO I

**INTRODUCCIÓN**

## Generalidades

En la presente tesis se busca reconocer el estado emocional de una persona a través de la voz, para así poder incrementar los índices de retención de los clientes que desean cancelar su tarjeta de crédito.

Como apoyo de información se ha usado un paper elaborado por Je-Bok Kim (University of Twente) y Jeong-Sik Park (Yeungnam University). Ellos han investigado sobre el Modelo de Redes Neuronales y han realizado una comparación con otros modelos, situando al MLLR como un modelo de alta efectividad.

Asimismo, también se ha usado el paper elaborado por Rosa González Hautamaki, Tomi Kinnunen, Ville Hautamaki y Anne María Laukkanen sobre la comparación de la verificación hecha por altavoces automáticos con verificación humana. Su principal objetivo es obtener información sobre la vulnerabilidad de los sistemas de verificación de altavoces y comparar su rendimiento con el de los oyentes humanos.

Ambos papers se complementan ya que uno de ellos explica la aplicación de un modelo en el reconocimiento de las emociones a través de la voz, mientras que el otro se centra en comparar la verificación del correcto reconocimiento por parte de los humanos y los compara con los altavoces automáticos.

Como ya se mencionó, el desconocimiento del estado emocional de los clientes por parte de las personas que atienden en un Call Center ha sido el origen de esta tesis ya que si los empleados del banco conocieran las emociones de los clientes aplicarían técnicas apropiadas y se incrementarían los índices de personas que aceptan no cancelar su tarjeta de crédito.

Es por ello que este tema tiene una alta importancia y su solución está basada en redes neuronales, y por otra parte también puede desarrollarse con distintos modelos.

Actualmente las empresas que cuentan con un Call Center no aplican ninguna técnica para reconocer las emociones de sus clientes; sin embargo, este tema ya se está estudiando en varios países para aplicarlo no solo en Call Centers sino también en otros ámbitos donde sería de mucha utilidad.

Para lograr el reconocimiento de las emociones se usará el lenguaje de programación Python con lo que haremos uso librerías basadas en Machine Learning.

Como ya se mencionó, actualmente ya no solo se limita a modelos o redes neuronales, sino que está yendo más allá al ser los datos de gran cantidad. Actualmente se aplica Deep Learning en el reconocimiento de la voz, reconocimiento de expresiones faciales, entre otros campos.

Si las empresas en el Perú se concentran en invertir sus esfuerzos y dinero en nuevas tecnologías haciendo uso de la inteligencia artificial, sus ventas aumentarían y con lo mismo sus ingresos, además que poco a poco se dejaría de utilizar recursos humanos con lo que se reducirían los gastos de este tipo y se automatizarían los procesos.

Por ello se deben realizar más investigaciones de las nuevas tecnologías para aplicarlas en empresas que le den un buen uso y así se satisfagan nuevas necesidades con reducción de tiempos y costos.

## 1.2. Problemática

Una empresa de banca en el Perú orienta su ofrecimiento de tarjeta de crédito a través de ciertos canales como Oficinas Bancarias, Canal Online, Call Center, entre otros.

Cada uno de estos canales se enfrenta a distintas situaciones en cuanto a la atención de los clientes.

Particularmente el canal Call Center debe resolver 3 situaciones:

* Dimensionamiento
* Tráfico
* Tiempo de atención

Estas 3 variables influyen considerablemente en los indicadores de ese canal.

* El dimensionamiento se refiere a la cantidad de empleados que están disponibles para atender las llamadas.
* El tráfico es la congestión en la línea telefónica.
* El tiempo de atención es el tiempo que un empleado utiliza para atender a un cliente.

Si hay muy pocos empleados y la cantidad de llamadas es grande, se genera tráfico, lo que ocasiona que un cliente deba esperar más tiempo de lo que debería para que se le responda la llamada. Una de las formas de que no se ocasione tráfico es reduciendo el tiempo de atención de los clientes. Como vemos, las 3 variables se relacionan.

El canal Call Center se divide a su vez en 2 rubros: Call In y Call Out. En el primero son los clientes quienes llaman, aquí lo que se busca es reducir el tiempo de atención.

En el caso de Call Out los vendedores son quienes llaman a los clientes para ofrecer sus productos. Evidentemente no se busca reducir el tiempo de atención, en este caso no es relevante cuánto dura la llamada, ya que el objetivo es que el cliente acepte la oferta. Tanto Call In o Call Out tienen un área de Retención.

Hay clientes de esta empresa que buscan dar de baja a su tarjeta, una de las maneras que utilizan para poder hacer el cambio es comunicándose con el personal de Call Center, para ello este canal cuenta con un área de Retención, donde los vendedores intentan que el cliente se mantenga como tal, para ello aplican técnicas comerciales.

El canal Call Center mide la eficiencia de estos vendedores a través de la Tasa de Efectividad de Retención. Evidentemente lo que se busca es incrementar esta tasa.

Uno de los indicadores del canal Call Center es la Tasa de Efectividad de Retención, con el que se mide cuántos de los clientes que han hecho una llamada quieren dar de baja su tarjeta de crédito.

Si los vendedores del Call Center conocieran cuál es el estado de ánimo del cliente al que buscan retener podrían aplicar mejores técnicas de retención. Y así, con menos clientes que hayan buscado dar de baja su tarjeta, la tasa de efectividad de retención aumentaría. En los Call Center se busca que esta tasa sea la máxima posible.

### **1.2.1. Formulación del problema**

* ¿El reconocimiento del estado emocional de una voz grabada en un Call Center de una entidad bancaria ubicada en San Isidro, con gestores de cobranza del mismo distrito, permite incrementar la tasa de retención de clientes, que buscan dar de baja a su tarjeta de crédito, durante el mes de diciembre del año 2018?

Datos generales del Grupo Poblacional, Gestores de Cobranza (GdC) que trabajan en la entidad bancaria ubicada en el distrito de San Isidro:

|  |  |
| --- | --- |
| **Población estimada** | 45 GdC |
| **Área geográfica** | Banco Falabella (San Isidro) |
| **Provincia** | Lima |
| **Departamento** | Lima |

**Tabla 1: Información de GdC de Banco Falabella, San Isidro**

*Fuente: Información extraída de Gestión Humana, Banco Falabella*

## 1.3. Objetivos

### 1.3.1. Objetivo General

Incrementar la tasa de retención de clientes de una entidad bancaria.

### 1.3.2. Objetivos Específicos

* Aplicar el Modelo de Redes Neuronales en el corpus de voces de las personas.

* Diferenciar las voces de las personas con las de cualquier otro ruido.

* Comparar esos resultados obtenidos del Modelo de Redes Neuronales con los obtenidos por una persona que tiene conocimiento en técnicas para reconocer los estados emocionales.

* Identificar cuántas voces sí fueron reconocidas por el Modelo de Redes Neuronales.

## 1.4. Hipótesis

### 1.4.1. Hipótesis General:

El reconocimiento de las emociones a través de la voz permitirá incrementar la tasa de retención de clientes.

### 1.4.2. Hipótesis Específicas:

* El MRN es factible de aplicar en el corpus de voces de las personas.
* El MRN ayudará a diferenciar la voz de las personas de los ruidos.
* La comparación de los resultados del MRN con los obtenidos por una persona arrojan un ratio de efectividad mayor al 40%.
* El MRN nos ayudará a cuantificar cuántas voces sí fueron correctamente reconocidas con su respectivo estado emocional.

## 1.5. Justificación

La presente tesis se ha elaborado con el fin de servir de apoyo a las empresas no solo del rubro bancario, sino para su aplicación en aquellas empresas que cuentan con un Call Center y que necesitan o bien captar más clientes o bien retener a los clientes ya existentes pero que quieren migrar a la competencia que con el pasar de los años es cada vez mayor para una organización.

# CAPÍTULO II

**FUNDAMENTO TEÓRICO**

## 2.1. Antecedentes

### **2.1.1. “Multistage data selection based unsupervised speaker adaptation for personalized speech emotion recognition”**

**Speech Emotion Recognition (SER):** Reconocimiento de las emociones a través del habla.

**Speaker Independent (SI):** Este enfoque construye modelos de emoción acústica utilizando datos de entrenamiento obtenidos de un grupo específico de hablantes que no son relevantes para usuarios reales. El enfoque SI es simple y eficaz para aplicaciones comunes, pero no siempre garantiza un rendimiento estable debido a las características acústicas inigualables entre los altavoces en los datos de entrenamiento y los usuarios reales.

**Speaker Dependent (SD):** El marco del modelo SD puede manejar eficientemente el problema de la variación entre altavoces, ya que los modelos acústicos están construidos sólo utilizando datos del usuario del sistema. Sin embargo, este enfoque tiene limitaciones significativas en las aplicaciones comerciales debido a la dificultad de recoger una cantidad suficiente de datos de emoción de los usuarios individuales.

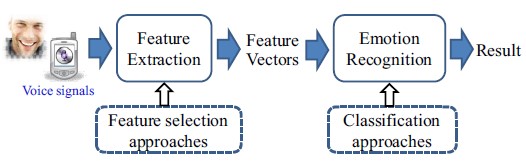
**Speaker Adapted (SA):** SA representa un modelo transformado a partir del modelo SI según los procedimientos de adaptación del hablante. La adaptación sólo requiere una cantidad relativamente pequeña de datos (llamados datos de adaptación) obtenidos del usuario (llamado el altavoz objetivo), pero produce el modelo acústico caracterizado por el usuario, casi logrando el rendimiento del modelo SD.

**Modelo Acústico Basado en el SER:**

* Resume el proceso estándar del SER que consiste en la extracción de vectores de características acústicas para poder identificar el estado emocional. Para ello utiliza un conjunto de características óptimas para representar el estado emocional del hablante.

* Para la clasificación se usan varios algoritmos basados en Machine Learning como:

* + Markov Model (HMM),
  + Gaussian Mixture Model (GMM),
  + Support Vector Machine (SVM)



**3.1.**

**Figura 1: Modelo Acústico Basado en el Ser**

*Fuente: Paper “Multistage data selection based unsupervised speaker adaptation for personalized speech emotion recognition”*

* Entre estos métodos, los clasificadores basados en modelos acústicos como GMM son para clasificar las emociones utilizando características acústicas a corto plazo como la energía.

* Para identificar el tipo de emoción de los enunciados de entrada, la probabilidad de cada GMM para un enunciado se calcula de la siguiente manera:

Donde:

Significa una secuencia de vectores característicos que son extraídos de los enunciados de entrada.

Y un GMM que indica el modelo acústico que le corresponda

Si hay E emociones

**MLLR-based speaker adaptation for SER:**

* Varias técnicas de adaptación, tales como Maximum Likelihood Linear Regression (MLLR) y Maximum a Posteriori (MAP) se han aplicado con éxito en las tareas de reconocimiento. SER tiene limitaciones en la capacidad de elección datos.

* En la figura se representa un procedimiento general para la adaptación convencional de MLLR. La adaptación de MLLR revisa los parámetros iniciales, es decir, las variables y variaciones gaussianas, así como las matrices de conversión. Para obtener la colección de datos de los oradores, las matrices de transformación son estimadas mediante el algoritmo Estimation-Maximization (EM).

### **2.1.2. “Emotion Recognition In Speech Signal”**

El artículo describe un estudio experimental sobre la expresión y el reconocimiento de las emociones vocales y el desarrollo de un agente informático para el reconocimiento de emociones. El estudio trata de un corpus de 700 expresiones breves que expresan cinco emociones: felicidad, ira, tristeza, miedo y estado normal (no emocional), que fueron retratadas por treinta personas. Los enunciados fueron evaluados por veintitrés personas, veinte de las cuales participaron en la grabación. La precisión de las emociones de reconocimiento en el habla es la siguiente: felicidad: 61.4%, ira: 72.2%, tristeza: 68.3%, miedo: 49.5% y normal: 66.3%. La capacidad humana para retratar emociones está aproximadamente en el mismo nivel (felicidad: 59.8%, enojo: 71.7%, tristeza: 68.1%, miedo: 49.7% y normal: 65.1%), pero la desviación estándar es mucho más grande. La capacidad humana para reconocer sus propias emociones también ha sido evaluada. Resultó que las personas son buenas en reconocer la ira (98.1%), la tristeza (80%) y el miedo (78.8%), pero tienen menos confianza en el estado normal (71.9%) y la felicidad (71.2%). Una parte del corpus se utilizó para extraer características y entrenar reconocimientos basados ​​en computadora. Se seleccionaron algunas estadísticas de la afinación, el primer y segundo morfema, la energía y la velocidad del habla y se crearon y compararon varios tipos de reconocedores. Los mejores resultados se obtuvieron utilizando los conjuntos de reconocedores de redes neuronales, que demostraron la siguiente precisión: estado normal: 55-75%, felicidad: 60-70%, enojo: 70-80%, tristeza: 75-85% y miedo. - 35-55%. La precisión promedio total es de aproximadamente 70%. Se creó un agente de reconocimiento de emociones que puede distinguir entre dos estados - "agitación" y "calma" - con la precisión del 77%. El agente se usó como parte de un sistema de soporte de decisiones para priorizar los mensajes de voz y asignar un agente humano adecuado para responder al mensaje en el entorno del centro de llamadas.

### **2.1.3. Detection and Analysis of Human Emotions through Voice and Speech Pattern Processing**

Este artículo propone un enfoque para la detección y análisis de emociones humanas a través de la voz. Ha sido desarrollado con el objetivo de incorporar sistemas de inteligencia artificial futuristas para mejorar la interacción hombre – ordenador.

Al inicio de la investigación, se definen conceptos como:

* Pitch: El pitch es una sensación auditiva en la que un oyente asigna tonos musicales a posiciones relativas en una escala musical basándose principalmente en su percepción de la frecuencia de vibración. El tono se puede cuantificar como una frecuencia, pero se basa en la percepción subjetiva de una onda de sonido. Las oscilaciones del sonido se pueden medir para obtener una frecuencia en hercios o ciclos por segundo. El tono es independiente de la intensidad o amplitud de la onda de sonido.
* Sonoridad: La sonoridad es una percepción subjetiva de la presión del sonido y se puede definir como el atributo de la sensación auditiva, en términos de los cuales, los sonidos se pueden ordenar en una escala que va desde silenciosa a fuerte. La presión acústica es la desviación de la presión local de la presión atmosférica ambiental, media o de equilibrio, causada por una onda acústica. El nivel de presión acústica (SPL) es una medida logarítmica de la presión efectiva de un sonido en relación con un valor de referencia y se mide a menudo en unidades de decibelios.

El límite inferior de audibilidad se define como SPL de 0 dB, pero el límite superior no está tan claramente definido.

* Timbre: Timbre es la calidad de sonido percibida de una nota musical, sonido o tono. Timbre distingue diferentes tipos de producción de sonido y permite a los oyentes distinguir diferentes instrumentos en la misma categoría. Las características físicas del sonido que determinan la percepción del timbre incluyen espectro y envolvente. En términos simples, timbre es lo que hace que un sonido en particular se perciba de manera diferente a otro sonido, incluso cuando tienen el mismo tono y volumen.

Principalmente se hace uso de este paper por las definiciones que plantea y por los rangos de frecuencia para la emoción de enojo y neutra que mencionan. Estos rangos se expresan en 2 tablas que en el Marco Teórico se explican con más detalle. (Ver tabla 1 y tabla 2).

## 2.2. Marco Teórico

Las emociones son parte esencial e inherente en la vida de las personas, además de que están presentes en el quehacer cotidiano.

Según la Psicología, una emoción es un estado afectivo que experimentamos, una reacción subjetiva al ambiente que viene acompañada de cambios orgánicos (fisiológicos y endocrinos) de origen innato, influidos por la experiencia.

Sin embargo, existe una confusión respecto a la diferencia entre una emoción y un estado de ánimo.

… Podríamos considerar que el estado de ánimo es una sensación de fondo que persiste en el tiempo. Normalmente, apenas percibimos nuestros estados de ánimo, pero, algunas veces, pueden llegar a ser muy intensos e insoportables. Los estados de ánimo no son lo mismo que las emociones, aunque comparten muchos aspectos en común. A veces, los estados de ánimo se caracterizan por ser de menor intensidad y mayor duración que las emociones…A diferencia de la mayoría de emociones, en los estados de ánimo no parece que exista una causa identificable…1

También, el estado de ánimo predispone a actuar de determinadas maneras. Un estado de ánimo determinado favorece que se incremente la probabilidad de que ocurra una determinada conducta.

…Los estados de ánimo pueden ser considerados como indicadores generales, tanto del funcionamiento fisiológico como de la experiencia psíquica. El estado de ánimo es algo así como un termómetro clínico que refleja todos los acontecimientos externos e internos que nos afectan...2

1. Robert E. Thayer, Nueva York, 1996, pág 22.
2. Robert E. Thayer, Nueva York, 1996, pág 20.

Continuando con las emociones, estas son de 6 tipos: miedo, asco, enfado, sorpresa, felicidad y tristeza. (Paul Eckman, 1972).

En la presente tesis se ha trabajado únicamente con 4 emociones: enojo, alegría, tristeza y neutralidad. Tres de estas 4 emociones han sido estudiadas por Paul Eckman, la última, neutralidad podría estar enmarcada dentro del modelo de Robert Plutchik en 1980, quien define la “rueda de emociones”, este modelo demuestra cómo las diferentes emociones pueden ser combinados o mezclados entre sí, del mismo modo en que un artista mezcla colores primarios para crear otros colores.

Como enojo se denomina el sentimiento desagradable que experimentamos cuando nos sentimos contrariados o atropellados por las palabras, las acciones o las actitudes de otros. La palabra, como tal, se deriva de la palabra “enojar”, que proviene del latín vulgar *inodiāre*, que significa ‘enfadar’.3

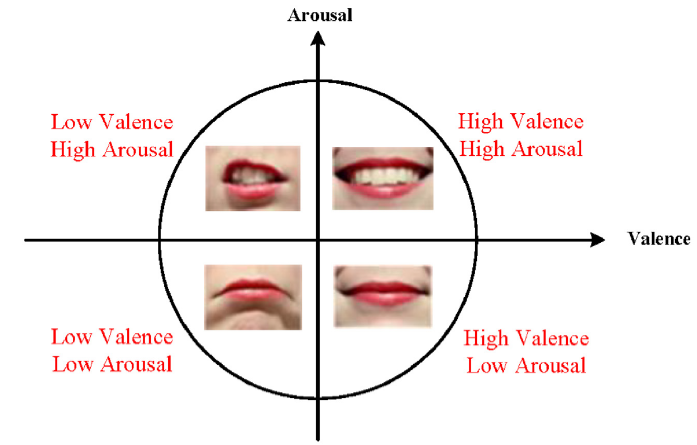
La felicidad es el estado emocional de una persona feliz; es la sensación de bienestar y realización que experimentamos cuando alcanzamos nuestras metas, deseos y propósitos; es un momento duradero de satisfacción, donde no hay necesidades que apremien, ni sufrimientos que atormenten.4

Tristeza es un estado anímico que ocurre por un acontecimiento desfavorable que suele manifestarse con signos exteriores como el llanto, pesimismo, melancolía, falta de ánimo, baja autoestima, en otros estados de insatisfacción. La palabra tristeza proviene del latín trístĭtĭa.5

La emoción neutral es la emoción más breve de todas cuantas hemos estudiado. Se produce de forma súbita ante una situación novedosa o extraña y desaparece con la misma rapidez con que apareció. Además, suele convertirse también rápidamente en otra emoción, la que sea congruente con la situación estimular desencadenante de la sorpresa.6

1. Significados.com. Disponible en: https://www.significados.com/enojo/Consultado: 27 de octubre de 2018, 12:08 am.
2. Significados.com. Disponible en: https://www.significados.com/felicidad/Consultado: 27 de octubre de 2018, 12:14 am.
3. Significados.com. Disponible en: https://www.significados.com/tristeza/Consultado: 27 de octubre de 2018, 12:14 am.
4. Psicología Online. Disponible en: https://www.psicologia-online.com/emociones-neutras-la-sorpresa-639.html/Consultado: 27 de octubre de 2018, 12:19 am.

Se utilizará la clasificación de las emociones propuesta por la Teoría Russel's Circumplex. Esta teoría indica que los estados emocionales o emociones, se distribuyen en un espacio circular bidimensional, con dimensiones de excitación y valencia. La excitación es el eje vertical y mide la intensidad de la activación emocional, mientras que la valencia es el eje horizontal y describe si la emoción es negativa o positiva. Podemos ver en la Figura 1 cómo se clasifican las emociones según esta teoría. Como podemos ver, la "Alta Valence, Low Arousal" está mapeada en el cuadrante inferior derecho, mientras que "Low Valence, High Arousal" está mapeada en el cuadrante superior izquierdo. La "Alta Valence, High Arousal" está en algún lugar en el cuadrante superior derecho, mientras que "Low Valence, Low Arousal" está en el cuadrante inferior izquierdo. Al determinar el valor de valencia de excitación de cierta emoción, podemos reconocer las emociones como las cuatro clases. De esta forma, la tarea de reconocimiento de emociones se divide en dos tareas de clasificación binarias.



**Figura 2: Teoría de Excitación y Valencia**

*Fuente: Paper “Respiration-based emotion recognition with deep learning”*

Asimismo, dentro de la investigación de la presente tesis, surge la interrogante de cuáles son las características medibles de una emoción, una de ellas, es la intensidad de la voz de una persona emocionada.

Dada la importancia de las emociones para la adaptación del comportamiento, uno puede asumir que la intensidad de los patrones de respuesta y la correspondiente experiencia emocional sea relativamente alta, lo que sugiere que esto puede ser un enfoque caracterológico importante al distinguir emociones de estados de ánimo, por ejemplo.7

Las siguientes tablas propuestas por Poorna Banerjee Dasgupta, muestran el tono medido en Hz para 2 estados emocionales de interés para la presente tesis:

Estado Neutro

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Pitch (Hz)** | **SPL (dB)** | **Timbre ascend time (s)** | **Timbre descend time (s)** | **Time gaps between words (s)** |
| **Speech Sample 1** | 1248 Hz | Gain -50 dB | 0.12 s | 0.11 s | 0.12 s |
| **Speech Sample 2** | 1355 Hz | Gain -48 dB | 0.06 s | 0.05 s | 0.12 s |

**Tabla 2: Características del Estado Neutro**

*Fuente: Paper* *“Detection and Analysis of Human Emotions*

*through Voice and Speech Pattern Processing”*

Estado Enojado

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Pitch (Hz)** | **SPL (dB)** | **Timbre ascend time (s)** | **Timbre descend time (s)** | **Time gaps between words (s)** |
| **Speech Sample 1** | 1541 Hz | Gain -30 dB | 0.13 s | 0.10 s | 0.09 s |
| **Speech Sample 2** | 1652 Hz | Gain -29 dB | 0.06 s | 0.04 s | 0.10 s |

**Tabla 3: Características del Estado Enojado**

*Fuente: Paper “Detection and Analysis of Human Emotions*

*through Voice and Speech Pattern Processing”*

7. Scherer, Klaus R., 2005, pág 702

En ambas tablas, la primera columna, Pitch, describe los niveles de intensidad medido en Hz para las 2 emociones descritas, Neutro y Enojado.

El portal web Recursos TIC menciona “cuanto mayor sea la amplitud de la onda, (A), mayor será su [intensidad](http://recursostic.educacion.es/secundaria/edad/2esobiologia/2quincena4/2q4_contenidos_2d1.htm). Por lo tanto, el sonido se oirá más fuerte”.8

Cabe resaltar la distinción entre intensidad y frecuencia, para esta tesis nos referimos a intensidad medida en Hertz, para de alguna manera tener una relación con la amplitud.

En las siguientes líneas, mencionaré los conceptos físicos como intensidad, amplitud y frecuencia para que las diferencias queden más claras.

La intensidad del sonido percibido, o propiedad que hace que éste se capte como fuerte o como débil, está relacionada con la intensidad de la onda sonora correspondiente, también llamada intensidad acústica. La intensidad acústica es una magnitud que da idea de la cantidad de energía que está fluyendo por el medio como consecuencia de la propagación de la onda.

Se define como la energía que atraviesa por segundo una superficie unidad dispuesta perpendicularmente a la dirección de propagación. Equivale a una potencia por unidad de superficie y se expresa en W/m2. La intensidad de una onda sonora es proporcional al cuadrado de su frecuencia y al cuadrado de su amplitud y disminuye con la distancia al foco.9

En el caso de las ondas sonoras, la amplitud de onda revela la distancia que existe entre el pico de la onda (el valor más alto) y su base, midiéndose en decibeles. A medida que crece la amplitud de onda, aumentan los decibeles, lo que refleja un crecimiento de la intensidad (el volumen) del [sonido](https://definicion.de/sonido).10

La frecuencia de un sonido se mide en Hercios (Hertz, Hz) y describe la cantidad de ondas por segundo que completan un ciclo. Es el tono o altura del sonido. Así se diferencia un sonido agudo de uno grave. El oído humano es capaz de percibir las frecuencias que se encuentren entre 20 Hertz y 20 kilohertz.11

1. Recursos TIC. Disponible en: http://recursostic.educacion.es/secundaria/edad/2esobiologia/2quincena4/2q4\_contenidos\_2d1.htm/Consultado: 20 de noviembre de 2018, 23:12 pm.
2. Sociedad de la Información. Disponible en: http://www.sociedadelainformacion.com/departfqtobarra/ondas/SONIDO/SONIDO.HTM/Consultado: 20 de noviembre de 2018, 23:12 pm.
3. Definicion.de. Disponible en: https://definicion.de/amplitud-de-onda//Consultado: 20 de noviembre de 2018, 23:12 pm.
4. Vix. Disponible en: https://www.vix.com/es/btg/curiosidades/2010/09/22/propiedades-del-sonido /20 de noviembre de 2018, 23:12 pm.

En la presente tesis, usaré cuatro categorías o tipos de emociones, que ya han sido definidas anteriormente:

|  |
| --- |
| CATEGORÍA DE EMOCIONES |
| Neutro |
| Feliz |
| Enojado |
| Triste |

**Tabla 4: Categoría de Emociones**

*Fuente: Elaboración propia*

Durante la comunicación oral la información de las emociones es implícita, así complementan la información explícita que se envía en el intercambio de un mensaje. Una de las maneras implícitas de transmitir una emoción mediante la comunicación es a través de la voz.

La voz humana es producida en la laringe, cuya parte esencial, la glotis, constituye el verdadero órgano de fonación humano. El aire procedente de los pulmones, es forzado durante la espiración a través de la glotis, haciendo vibrar los dos pares de cuerdas vocales, que se asemejan a dos lengüetas dobles membranáceas. Las cavidades de la cabeza, relacionadas con el sistema respiratorio y nasofaríngeo, actúan como resonadores.

El aparato de fonación puede ser controlado conscientemente por quien habla o canta. La variación de la intensidad depende de la fuerza de la espiración. En el hombre las cuerdas vocales son algo más largas y más gruesas que en la mujer y el niño, por lo que produce sonidos más graves. La extensión de las voces es aproximadamente de dos octavas para cada voz.

Uno de los primeros intentos científicos sobre el estudio de la entonación fue el que realizó, en el ***año 1774, el inglés James Burnett*** que estaba muy preocupado por la existencia de entonación en su idioma. Sostenía que en el inglés no había entonación y así lo expresó en sus escritos diciendo que “la música en nuestro idioma no es más que la música de un tambor”.  
  
***Un año más tarde, su compatriota Josué Steele*** publicó la obra titulada Ensayo dedicado a establecer la melodía y la medida del habla (Essay towards establishing the melody) sobre la transcripción de la entonación y en ella rechazaba esa afirmación de un modo muy original. Colocó una tira de papel en el mástil de una viola, marcó con lápiz los lugares que correspondían a las diferentes notas y después fue imitando con la viola las inflexiones de la voz. Se dio cuenta de que los tonos del habla se deslizaban sobre varias notas e ideó entonces un sistema notacional que tomaba en consideración esta peculiaridad: fue el primer método sistemático de transcripción del inglés.  
  
Posteriormente, se han realizado estudios sobre transcripciones contemporáneas de la entonación que han contribuido a ampliar enormemente el conocimiento de este tema ya que reflejan diferentes puntos de vista sobre su naturaleza.  
  
También se han llevado a cabo diversos estudios intentando establecer relaciones entre tipos corporales y voces, habiéndose observado ciertas correlaciones entre ambos aunque con un amplio margen de error. Sin embargo, es indudable la relación entre determinadas características físicas y la voz. Así, está el fenómeno del cambio de voz que acompaña el desarrollo de las características sexuales secundarias durante la pubertad, debido al rápido crecimiento de la laringe (1 cm., en los chicos y 0,3 a 0,4 cm., en las chicas) que da lugar a una mayor extensión de aproximadamente una octava musical más baja en los chicos. En la vejez las voces se hacen más graves, ásperas y temblorosas, debido a la disminución de la eficiencia de los órganos vocales y de la respiración.  
  
Igual que tenemos las huellas digitales también poseemos las huellas vocales. La identificación de personas mediante sus huellas vocales (sonograma) ha sido utilizada en tribunales de EE.UU. y, aunque sigue siendo una técnica controvertida, su potencial en el diagnóstico médico es muy alto sobre todo con el uso de los ordenadores para distinguir ruidos anormales en la emisión de la voz. De esta forma, se pueden identificar sonogramas con frecuencias muy melódicas y otros con frecuencias mucho más irregulares.  
  
Finalmente, en cuanto a la intensidad de la emoción, si ésta es fuerte, el ritmo respiratorio pierde regularidad y los órganos que intervienen en la articulación de la voz alteran su funcionamiento. Dicha alteración no refleja un tipo particular de sentimiento sino el grado de intensidad del sentimiento. Este proceso comienza con la excitación producida por los movimientos convulsos en el diafragma, lo cual interrumpe con sacudidas los movimientos de la laringe y resulta ser la voz que llamamos “entrecortada o balbuciente”.  
  
Un grado más es el que produce una fuerte contracción en la laringe de tal modo que las cuerdas vocales apenas pueden vibrar, lo que ofrece un resultado de voz “enronquecida o ahogada”. Cuando la emoción es muy fuerte, puede producirse una paralización momentánea de la corriente respiratoria y es lo que solemos describir como “un nudo en la garganta”.

# CAPÍTULO III

**MODELO DE SOLUCIÓN**

1. Extracción de la muestra:

Datos reales:

* Tamaño de la población (N): 45 GdC

Datos asumidos:

* Nivel de confianza (95%): Z = 1.96
* Probabilidad de éxito (p) = 0.50
* Probabilidad de fracaso (q) = 0.50
* Error máximo admisible (d) = 0.04

Fórmula estadística a utilizar:

Aplicando la fórmula con los datos reales y con los datos asumidos, se obtiene:

**GdC**

1. Se identifica el grupo Control y el grupo Experimental.

* Grupo Control: El 40% de muestra conforma el grupo control; es decir, 16 GdC. Este grupo escuchará cada audio de voz y los relacionará con la emoción que creen que corresponde, las cuales pueden ser una emoción de alegría, enojo, tristeza o neutra.
* Grupo Experimental: Está compuesto por la metodología en base a redes neuronales, que reconoce cada emoción según la frecuencia o pitch del sonido de la voz.

1. La siguiente tabla muestra los audios con su respectiva emoción, la base de datos es de DataSet Reddy. Los rostros que aparecen en las imágenes son referenciales puesto que, en la data original, eran videos mp4, que han sido convertidos a audios en formato wav.

Base de Datos

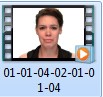
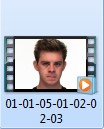
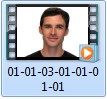
Tipo de Emoción

Feliz

Enojado

Triste

Neutro

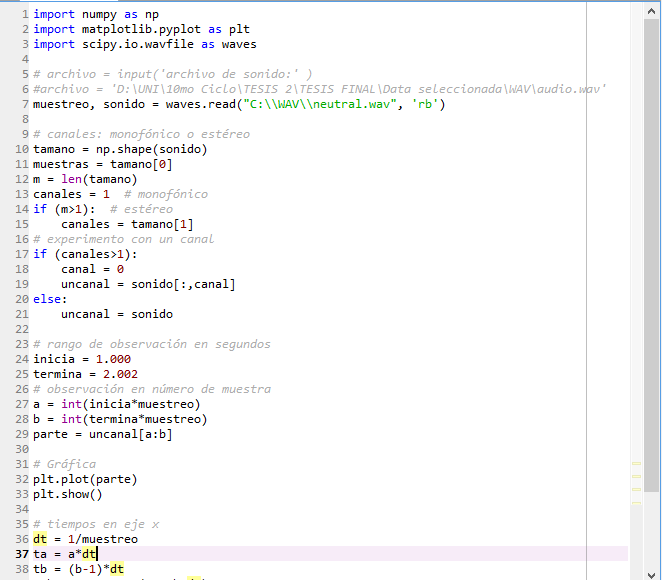


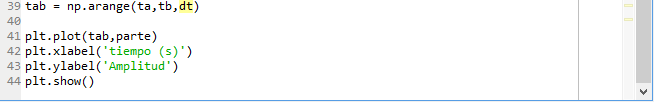
***Tabla 5: Base de Datos de Emociones***

*Fuente: Elaboración Propia*

1. El siguiente paso es transformar esa data a formato WAV, ya que se encuentra como MP4.
2. Se calcula la amplitud en función del tiempo para cada audio de voz.

Para ello usamos el lenguaje de programación Python:



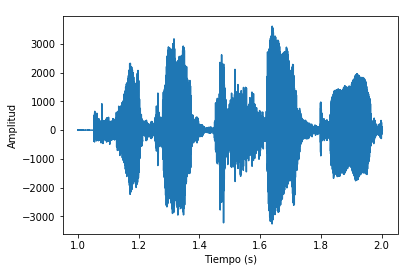


**Figura 3: Imagen de código en Python**

*Fuente: Elaboración propia*

Las siguientes imágenes muestran para cada emoción, la gráfica en función del tiempo y la amplitud, luego de ejecutar el programa ya mencionado:

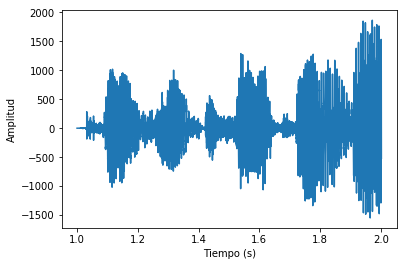
* Triste:



**Figura 4: Emoción Triste**

*Fuente: Elaboración propia*

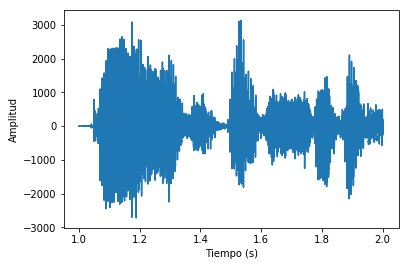
* Alegre:



***Figura 5: Emoción Alegre***

*Fuente: Elaboración propia*

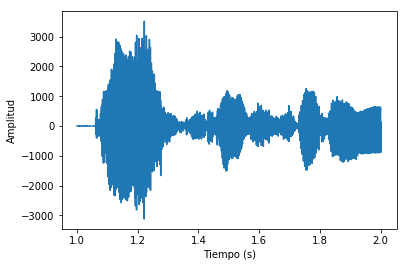
* Enojado:



**Figura 6: Emoción Enojado**

*Fuente: Elaboración propia*

* Neutro:

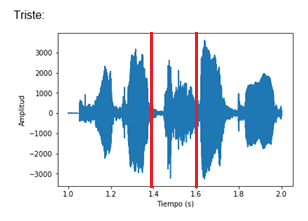
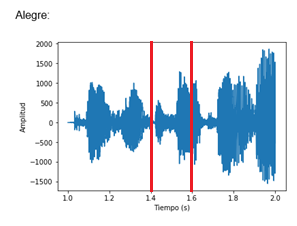


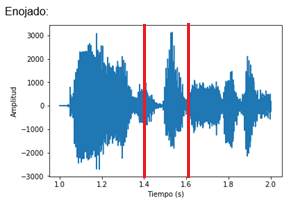
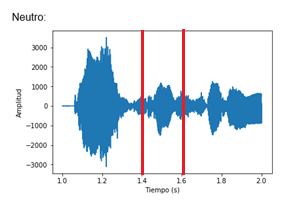
**Figura 7: Emoción Neutro**

*Fuente: Elaboración Propia*

Esto se hace con el fin de que una vez obtenida la amplitud en función del tiempo se pueda calcular la frecuencia.

1. Una vez que se tiene la emoción con su respectiva gráfica, se elige arbitrariamente un rango de tiempo a analizar, en este caso se tomará el rango de tiempo de [1.4 – 1.6] s.





**Figura 8: Rango de tiempo para las cuatro emociones**

*Fuente: Elaboración propia*

1. Cada emoción tiene una determinada amplitud en el segmento escogido, que es del 1.4 a 1.6 s y considerando la amplitud desde 0.
2. Como se explicó en el marco teórico, la amplitud está ligada con el pitch de una manera directa.

Si recordamos las 2 tablas propuestas por Poorna Banerjee Dasgupta:

Estado Neutro

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Pitch (Hz)** | **SPL (dB)** | **Timbre ascend time (s)** | **Timbre descend time (s)** | **Time gaps between words (s)** |
| **Speech Sample 1** | 1248 Hz | Gain -50 dB | 0.12 s | 0.11 s | 0.12 s |
| **Speech Sample 2** | 1355 Hz | Gain -48 dB | 0.06 s | 0.05 s | 0.12 s |

**Figura 9: Pitch del Estado Neutro**

*Fuente: Paper propuesto por Poorna Banerjee Dasgupta*

Estado Enojado

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Pitch (Hz)** | **SPL (dB)** | **Timbre ascend time (s)** | **Timbre descend time (s)** | **Time gaps between words (s)** |
| **Speech Sample 1** | 1541 Hz | Gain -30 dB | 0.13 s | 0.10 s | 0.09 s |
| **Speech Sample 2** | 1652 Hz | Gain -29 dB | 0.06 s | 0.04 s | 0.10 s |

**Figura 10: Pitch del Estado Enojo**

*Fuente: Paper propuesto por Poorna Banerjee Dasgupta*

Notamos que el pitch de la emoción de enojo es superior al pitch de la emoción neutra, para cada uno de los speech sample.

En las muestras extraídas, contamos con la amplitud, que según se explicó en el marco teórico es directamente proporcional al pitch o intensidad.

1. Ahora estableceremos un rango para estas 2 emociones (enojo y neutra).

La regla es la siguiente:

* Si el pitch es MAYOR a 1541 Hz y MENOR a 1652 Hz, entonces le corresponde una emoción de enojo.
* Si el pitch es MAYOR a 1248 Hz y MENOR a 1355 Hz, entonces le corresponde una emoción neutra.

Esta regla está basada en el paper “Detection and Analysis of Human Emotions

through Voice and Speech Pattern Processing” elaborado por Poorna Banerjee Dasgupta. En los antecedentes, se explica a detalle dicha investigación.

1. El siguiente paso es utilizar una relación entre el pitch y la amplitud, ya que por las muestras tenemos amplitud, mas no pitchs.

La ecuación física nos dice:

Donde:

Ao: Amplitud inicial (db)

A: Amplitud máxima (db)

f: Frecuencia (Hz)

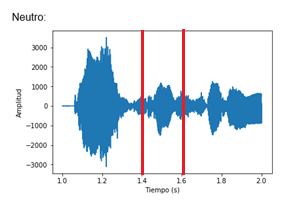
t: Tiempo (s)

ϕ: Fase inicial

Como se observa en las gráficas, la amplitud inicial es cero, y la fase inicial también, por lo que la fórmula quedaría así:

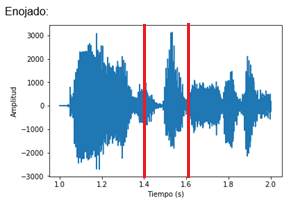
Ahora, se va a calcular la frecuencia o pitch (Hz) para cada emoción en dos tiempos (t = 1.4 s y t = 1.6 s)

* Para la emoción neutra:



Utilizando la ecuación anterior, se obtiene que el pitch es 1312 Hz.

* Para la emoción enojo:



Utilizando la ecuación anterior, se obtiene que el pitch es 1559 Hz.

1. Como se mencionó en el primer punto, existen dos grupos, el Grupo Control y el Grupo Experimental.

A continuación, se presenta una Matriz de Confusión, con los resultados obtenidos, una para cada grupo:

* Matriz de Confusión del Grupo Control, compuesto por 16 GdC

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Categoría** | **Triste** | **Alegre** | **Enojado** | **Neutro** |
| **Triste** | |||||||||||||||| (16) |  |  |  |
| **Alegre** |  | |||||||||| (10) |  | |||||| (6) |
| **Enojado** |  |  | |||| (4) | |||||||||||| (12) |
| **Neutro** |  | ||| (3) |  | ||||||||||||| (13) |

**Tabla 6: Matriz de confusión del Grupo Control**

*Fuente: Elaboración propia basada en paper Emotion Recognition in Speech Signal*

* Para el Grupo Experimental, con 1 iteración.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Categoría** | **Triste** | **Alegre** | **Enojado** | **Neutro** |
| **Triste** |  |  |  |  |
| **Alegre** |  |  |  |  |
| **Enojado** |  |  | | (1) |  |
| **Neutro** |  |  |  | | (1) |

**Tabla 7: Matriz de confusión del Grupo Experimental**

*Fuente: Elaboración propia basada en paper Emotion Recognition in Speech Signal*

En el capítulo siguiente se mostrará un resumen de los resultados del modelo de la solución.

# CAPÍTULO IV

**RESULTADOS**

En el capítulo anterior vimos los resultados de la Matriz de Confusión del grupo control y del grupo experimental, en base a esas tablas (tabla 6 y 7), se hará un cálculo de efectividades.

* Para el Grupo Control:

|  |  |
| --- | --- |
| **Categoría** | **Efectividad** |
| **Triste** | 100% |
| **Alegre** | 62.50% |
| **Enojado** | 25% |
| **Neutro** | 81.25% |

**Tabla 8: Resultados de Efectividad del Grupo Control**

*Fuente: Elaboración propia basada en Tabla 6*

* Para el Grupo Experimental:

|  |  |
| --- | --- |
| **Categoría** | **Efectividad** |
| **Triste** | - |
| **Alegre** | - |
| **Enojado** | 100% |
| **Neutro** | 100% |

**Tabla 9: Resultados de Efectividad del Grupo Experimental**

*Fuente: Elaboración propia basada en Tabla 7*

**RESUMEN DE RESULTADOS:**

* Para el Grupo Control:

|  |  |  |
| --- | --- | --- |
| **Categoría** | **Aciertos** | **Errores** |
| **Triste** | 16 | 0 |
| **Alegre** | 10 | 6 |
| **Enojado** | 4 | 12 |
| **Neutro** | 13 | 3 |

**Tabla 10: Resultados de aciertos y errores del Grupo Control**

*Fuente: Elaboración propia basada en Tabla 6*

* Para el Grupo Experimental:

|  |  |  |
| --- | --- | --- |
| **Categoría** | **Aciertos** | **Errores** |
| **Triste** | - | - |
| **Alegre** | - | - |
| **Enojado** | 1 | 0 |
| **Neutro** | 1 | 0 |

**Tabla 11: Resultados de aciertos y errores del Grupo Experimental**

*Fuente: Elaboración propia basada en Tabla 7*

En el caso del grupo experimental no hay referencias en cuanto a pitch para las emociones alegre y triste, es por ello que ambas no tienen datos.

Como puede notarse, en el grupo experimental se alcanza una mayor efectividad en la emoción de enojo, una efectividad al 100%. Un dato opuesto al que se obtiene del grupo control, donde doce personas no fueron capaces de realizar un correcto reconocimiento.

En el caso de la emoción alegre, el grupo control tuvo una efectividad del 62.5%. Lamentablemente aún no es posible compararlo con los resultados del grupo experimental porque aún no se tienen referencias de pitch de esta emoción.

Un dato bastante curioso, es que las personas que escucharon los audios, solían estar indecisos con la emoción neutra. En cada audio que escuchaban creían que se trataba de esta emoción. Solo en la emoción triste estaban totalmente seguros de su decisión.

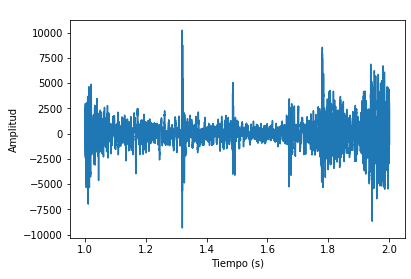
Lo que hace pensar que es muy peligroso que sea justamente la emoción enojo la que no es bien reconocida por una persona; porque por lo menos para el caso de una entidad bancaria que busca retener a sus clientes que desean cancelar su tarjeta de crédito, les puede ocasionar problemas y no lograr el objetivo de recapturar al cliente.

Finalmente, quisiera acotar que esta investigación busca ser el inicio de un estudio mucho más profundo en cuanto al reconocimiento de la emoción a través de la voz como apoyo para una entidad bancaria. Porque son muchos los problemas que puede enfrentar un banco en distintos ámbitos y un reconocimiento emocional, además, estrecharía los lazos que tiene con sus clientes y facilitaría la comunicación entre ellos con la empresa.

# ANEXOS

## Resultados del audio extraído por una persona fuera de la muestra

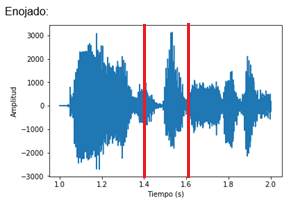
* Esta es la imagen de la función amplitud y tiempo del audio:



**Figura 11: Amplitud de audio de voz**

*Fuente: Elaboración propia*

* Para el rango de tiempo 1.4 y 1.6, se observa que la amplitud es de 5 000. Se compara esa amplitud con la gráfica siguiente, que representa la emoción enojo de una muestra extraída, sabiendo que la emoción efectivamente corresponde al enojo.



**Figura 12: Emoción enojo de muestra extraída**

*Fuente: Elaboración propia*

* Se observa que alcanza una amplitud superior a 3 000, lo que no hacen las demás emociones analizadas en esta tesis, por lo que se interpreta que la emoción a la que corresponde el último audio de este anexo es de enojo, con una amplitud de 5 000.

# BIBLIOGRAFÍA

Blog de Psicología. <https://www.psicoactiva.com/blog/que-son-las-emociones/>

Diccionario en línea. <https://www.significados.com/>

Jae-Bok Kim, Jeong-SikPark. Multistage data selection-based unsupervised speaker adaptation for personalized speech emotion recognition. ELSEVIER

Julián David Echeverry Correa, Mauricio Morales Pérez. Reconocimiento de emociones en el Habla.

Humberto Pérez Espinosa, Carlos Alberto Reyes García. Reconocimiento de Emociones a Partir de Voz Basado en un Modelo Emocional Continuo.

Coordinación de Ciencias Computacionales INAOE.

Poorna Banerjee Dasgupta. Detection and Analysis of Human Emotions

through Voice and Speech Pattern Processing. International Journal of Computer Trends and Technology (IJCTT)

Robert E. Thayer. El origen de los estados de ánimo cotidianos.

Rosa Gonzalez Hautamaki, Tomi Kinnunen, Ville Hautamaki, AnneMaria Laukkanen. Automatic versus human speaker verification: The case of voice mimicry. ScienceDirect