Índice

1. Contexto y Objetivo **2**

2. Fundamentos teóricos y revisión bibliográfica **3**
2.1. La voz . . . 3
2.1.1. Mecanismo de producción de voz . . . 3
2.1.2. Modelización de la señal de voz . . . 5
2.1.3. Procesamiento de la señal de voz . . . 8
2.2. Modelos de emociones . . . 9
2.2.1. Modelo discreto . . . 9
2.2.2. Modelo continuo . . . 9
2.3. Características de la señal de voz . . . 10
2.3.1. Prosodia . . . 10
2.3.2. Características espectrales . . . 14
2.3.3. Calidad de voz . . . 17
2.4. Selección de características y procesado de características . . . 18
2.5. Test de significación estadística . . . 20

3. Procesamiento de la señal y extracción de características **22**

4. Análisis de las señales de voz **24**
4.1. Grabación con alegría . . . 24
4.2. Primera grabación con enfado . . . 27
4.3. Segunda grabación con enfado . . . 30

5. Clasificación y Resultados **32**

6. Conclusiones y Trabajo Futuro **33**

Bibliografa **36**

A. Apéndice: Reuniones y Actividades **38**

1. Contexto y Objetivo

Xupera es una empresa con sede en Baracaldo orientada a dar servicio a grandes clientes entre los que se encuentran organizaciones empresariales como Euskaltel y Fagor entre otros y administraciones públicas como el Gobierno Vasco o el Gobierno del Principado de Asturias.
Esta empresa ofrece una gran variedad de servicios orientados a satisfacer las necesidades del cliente, la auditoría ante distintas situaciones, el diseño de nuevas estrategias o aplicaciones para Smatphone y Tablet, análisis de la reputación online, gestión de operaciones de procesos de venta, de atención al cliente o de fidelización y un largo etcétera. Aunque el enfoque principal es el servicio al cliente, Xupera cuenta con diferentes líneas de I+D+i que buscan la mejora de los procesos y servicios que ofrecen a sus clientes. Una de estas líneas pretende mejorar la calidad del servicio *call-center* buscando una respuesta tecnológica capaz de detectar las emociones del llamante.
Con este propósito la empresa contactó con el grupo de investigación Pattern Recognition and Speech Technology de la Universidad del País Vasco.

El objetivo de este proyecto de fin de grado es hacer una primera toma de contacto con la detección automática de emociones en el habla para que después se pueda desarrollar un proyecto capaz de dar una solución tecnológica eficaz a este requerimiento.
Para ello hicimos una reunión con la empresa con el fin concretar sus necesidades, ellos buscaban un algoritmo capaz de diferenciar la ira y la alegría en sus llamantes para así por una parte ofrecer un mejor servicio en llamadas realizadas por un autómata y por otra para poder evaluar más eficientemente la calidad de su servicio.

Este trabajo de fin de grado busca hallar las herramientas adecuadas dentro de la tecnología y de la física para dar respuesta a la demanda de una empresa. Para completar este objetivo se debe realizar lo siguiente:

- Comprender los fundamentos teóricos de la producción y análisis de la señal de voz.
- Conocer el estado del arte actual en el ámbito de la detección automática de emociones en el habla para así identificar los factores más relevantes a la hora de clasificar las emociones.
- Realizar el análisis de tres llamadas reales seleccionadas por la empresa en las que en dos se muestra enfado y en una alegría. Este análisis constará de dos partes, la primera se realizará en este trabajo y será una evaluación inicial de los fragmentos más significativos de las llamadas para ver si en primera instancia las características escogidas son las acertadas. La segunda la hará un alumno de matemáticas para su proyecto de fin de grado, creando un clasificador capaz de detectar la ira, la alegría y el estado neutro basándose en las características extraídas.

La estructura de la memoria pretende seguir el hilo temporal de la realización del proyecto. Por ello comenzaremos con la sección del fundamentos teóricos y revisión bibliográfica corresponderá al primer periodo del trabajo donde se hizo una labor bibliográfica. En esta parte se explicarán en profundidad distintos aspectos tanto de la voz como de las emociones creando la base teórica necesaria para el desarrollo del proyecto. A continuación se explicará con detalle el procesamiento y extracción de características de las tres muestras en vistas a los objetivos que queremos que cumpla nuestros clasificadores. Seguidamente se hará un análisis cualitativo del comportamiento de estas tres muestras uniéndolo con la teoría vista

en la primera parte del trabajo y un pequeño examen estadístico de los fragmentos más marcados. Aprovecharemos tamién para comentar brevemente los diferentes clasificadores que se usarán en el trabajo de fin de grado de matemáticas y los resultados obtenido. Por último se hará una valoración tanto de los resultados obtenidos en este trabajo como de los obtenidos por mi compañero.

2. Fundamentos teóricos y revisión bibliográfica

La detección de emociones en el habla es una disciplina relativamente novedosa y en continua evolución, por ello es importante la recolección de información de distintas fuentes y así evaluar el método que mejor se adecue a nuestro objetivo.
Se estudiarán los siguientes apartados:

- La voz
- Procesamiento de la señal de voz
- Modelos de emociones
- Características de la voz
- Selección de las características

2.1. La voz

2.1.1. Mecanismo de producción de voz

El uso de la voz con fin comunicativo es uno de los recursos más eficaces que tenemos el ser humano. La producción de voz es un mecanismo complicado que debe de ser entendido para poder crear un modelo a partir del cual se podrá estudiar su la variación dependiendo de las emociones que muestre el individuo.

La voz es un onda de presión acústica producida por los movimientos voluntarios del aparato respiratorio y masticatorio. Estos aparatos están constituidos por los pulmones, encargados de la inhalación llevada acabo por la expansión de la caja torácica, acción que disminuye la presión del aire en los pulmones haciendo que se introduzca aire en los pulmones. El tracto vocal es un tubo acústico no-uniforme que se extiende desde el final de la tráquea hasta los labios y el tracto nasal que empieza en el velo del paladar y termina en los orificios nasales.
El velo del paladar es el encargado de regular el canal de emisión del sonido. Para que el sonido sea nasal (n, m por ejemplo)el frente del tracto vocal se cierra totalmente y el velo se abre haciendo que la mayor parte de la transmisión del sonido sea por este canal. El sonido también puede ser vocal, como es el caso de las vocales en las cuales la participación del tracto nasal como aparato fonador es despreciable. Puede ocurrir que el sonido sea mixto como es el caso de la mayoría de consonantes y en especial de las consonantes fricativas[1] como la z o la f.

[1]las fricativas son producidas por un excitación incoherente del tracto vocal, este ruido es generado por un flujo de aire turbulento. Dependiendo del lugar de la producción del mismo y de la ausencia o presencia de voz se crearan las diferentes consonantes.

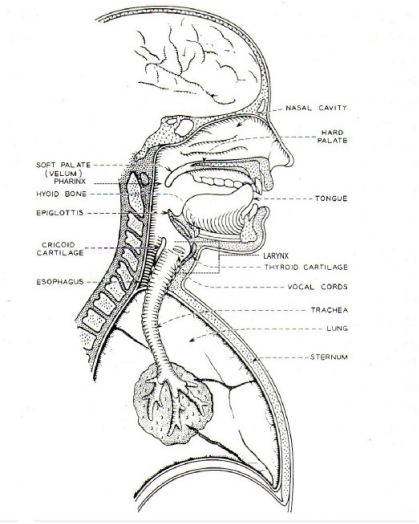

Figura 1. *Diagrama esquemático del mecanismo vocal humano, extraído de [3]*

Los músculos torácicos y abdominales son los que determinan la energía de la voz. Al contraerse aumentan la presión pulmonar. La frecuencia fundamental o pitch a la que vibren las cuerdas vocales será determinada por esta presión. Durante el discurso la presión pulmonar se mantiene relativamente constante por una contracción lenta y constante de la caja torácica. El aire exhalado de los pulmones fluye por la traquea hasta la faringe pasando por la laringe donde se encuentran las cuerdas vocales. La fonación o el sonido vocal está producido por la vibración de las mismas. Para que ocurra esta vibración la presión subglotal aumenta lo suficiente para separarlas por aceleración normal. Partiendo de una situación donde las cuerdas vocales están juntas, el principio de Bernoulli[2] hace que la velocidad del aire haga disminuir la presión y que por tanto las cuerdas se aproximen de nuevo. Al aproximarse la abertura entre las cuerdas vocales disminuye aumentando la presión subglotal y haciendo que se vuelvan a separar. El ciclo continuará mientras que continue el flujo de aire. El periodo al que vibran las cuerdas vocales lo determinan su masa, su flexibilidad y la presión subglotal. Como hemos mencionado a la frecuencia se le llama frecuencia fundamental o pitch, aunque varía con la presión (El pitch disminuye cuando la presión es baja), es una característica propia de cada individuo puesto que la masa y la elasticidad de las cuerdas vocales varia de un individuo a otro.
El volumen del flujo de aire que atraviesa la glotis por unidad de tiempo es aproximadamente proporcional al área de apertura glotal, en general la forma de la onda similar a una onda triangular. La transformada de fourier de una onda triangular es $\frac{\sin^2(x)}{x^2}$.E esta es la razón por la que la onda de voz sea rica en armónicos. La velocidad del volumen del flujo de aire corresponde a la *energía acústica.*

[2] $\frac{V^2\rho}{2} + P + \rho g z = constante$ donde V es la velocidad del fluido en la sección, ρ la densidad del fluido, P la presión, g la aceleración gravitatoria y z la altura

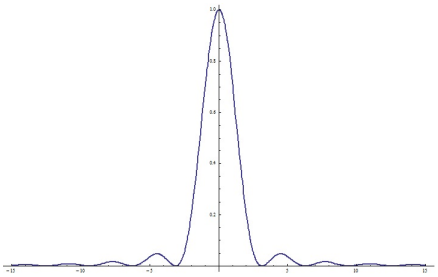

Figura 2. *Forma de la función* $sinc^2(x)$

La producción de voz tiene tres grandes características:

- Fuente sonora/ sorda: Relacionado con la presencia o ausencia de la vibración de las cuerdas vocales.
- Modo de articulación: Variando el modo de articular se producen distintos sonidos: Fricativo cuando se deja pasar el aire, oclusivo cuando se cierra el tracto vocal por un pequeño lapso de tiempo y etcétera
- Punto de articulación: Donde se produce el sonido.

Los sonidos continuantes sordos como el de la letra s, llamados así porque se producen sin que haya interrupción completa del flujo de aire y sin vibración de las cuerdas vocales, ocurren por otro tipo de excitación vocal. Estos sonidos los originan un turbulento flujo de aire producido por una contracción en algún punto del tracto, crea un sonido acústico y consecuentemente una excitación incoherente del sistema vocal.

Por otra parte los sonidos oclusivos como el de la b en el caso de los sonoros o la p en el de los sordos, ocurren por una excitación transitoria del tracto vocal al haber una liberación abrupta de la presión. En una primera aproximación la variación de la presión viene representada por una función de heaviside (o función escalón)[3], por lo tanto se puede esperar, por su transformada de Fourier, que el espectro sea proporcionalmente inverso a la frecuencia.

2.1.2. Modelización de la señal de voz

La voz es una onda sonora longitudinal de presión caracterizada por la frecuencia ω y el número de onda k siendo $k = \frac{2\pi}{\lambda}$, la frecuencia de la onda sonora va desde 20Hz hasta los 20KHz.
Hacer un modelo adecuado de la producción de voz es crucial para su análisis. El mecanismo de producción de voz es el siguiente. La fuente de alimentación está compuesta por los pulmones y los músculos encargados de la respiración, esto excita dos tipos de sonidos los sonoros y los sordos. Los sonoros se originan siempre por la vibración de las cuerdas vocales y su emisión puede ser vocal, nasal o mixta. Los sonidos sordos por su parte, se crean de dos formas, bien por el paso de una corriente de aire a través de una contracción en alguna parte del tracto, produciendo una turbulencia y un sonido incoherente, o bien por la liberación espontánea de la presión creada tras el cierre de la zona del tracto donde se origina, creando en este caso una excitación breve y transitoria. Los sonidos sordos son generalmente emitidos

[3] $u(t) = \begin{cases} 1 & t > 0 \\ 0 & t < 0 \end{cases}$ y su transformada de Fourier es $F[u(t)] = \frac{1}{i\omega}$

por la boca. Cada sonido refleja las características de su modo y lugar de origen así como del canal usado para su transmisión. (Ver Figura 3)

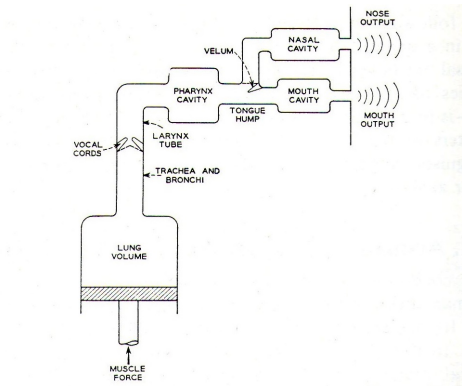

Figura 3. *Diagrama esquemático de los componentes funcionales del tracto vocal, extraído de [3]*

Uno de los modelos más usados por ser relativamente simple y dar buenos resultados [2], [3] y [4] es el modelo de predicción lineal. Comenzaremos explicando brevemente la teoría de los sistemas lineales para luego explicar el modelo usado en el análisis de la voz.
Un sistema lineal es aquel cuya salida es el resultado de la combinación lineal de la entrada actual y sus entradas previas. En el caso de que sus parámetros no varíen con respecto al tiempo la expresión matemática es la siguiente [10]:

$$y(n) = \sum_{j=0}^{q} b_j x(n-j) - \sum_{k=1}^{p} a_k y(n-k) \tag{1}$$

donde b_j y a_k representan los coeficientes de cada valor de entrada y salida respectivamente, n denota el número del elemento de la señal discreta y p y q el orden del sistema.
Haciendo una *transformada en Z*[4] se obtiene el siguiente resultado:

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{j=0}^{q} b_j z^{-j}}{\sum_{k=0}^{p} a_k z^{-k}} \tag{2}$$

H(z) es la función de transferencia y representa el cociente de la transformada de Z de la salida Y(z) y la transformada de z de la entrada X(z).

Existen en general tres tipos de funciones de transferencia que predicen el comportamiento de la función Y(z):

- La función del númerador es constante, dando lugar a una función todo polos.
- La función del denominador es constante, dando lugar a una función todo ceros.
- El caso más general donde no se puede asumir ninguna de las dos afirmaciones anteriores.

[4]La transformada en Z se usa para transformar una señal discreta en una función de la variable compleja z. $X(z) = \sum_{n=-\infty}^{\infty} x[n] z^{-n}$ donde n es un número entero y z un número complejo definido por $z = Ae^{j\omega}$, siendo A el módulo y ω la frecuencia en rad/s

El análisis de la señal de voz se puede hacer de una manera eficiente mediante un sistema autorregresivo (AR) o de todo polos. Esto significa que la salida depende linealmente de sus valores anteriores. La predicción lineal de todo polos es la más estudiada y la que veremos a continuación.

La predicción lineal de todo polos puede representarse de la siguiente manera:

$$y'(n) = -\sum_{k=1}^{p} a_k y(n-k) \tag{3}$$

y' representa la estimación del elemento n de la señal de salida y a orden de estimación p. El error reside en la diferencia de la señal *y(n)* y su estimación:

$$e(n) = y(n) - y'(n) \tag{4}$$

Tras definir la señal de error se usan diferentes métodos para predecir los coeficientes. Profundizar en este tema se escapa del objetivo de esta sección que no es otro que describir de una forma general la modelización de la voz comúnmente usada hoy en día para así tener una visión más completa de conceptos que se introducirán en futuras secciones. Por lo tanto en resumen diremos que principalmente se usan dos métodos, el método de autocorrelación que minimiza la señal de error en todo el espacio temporal y el método de covarianza que al contrario que el método anterior lo que minimiza es la suma total de la señal de error a lo largo del tramo deseado.

Como hemos visto hay dos tipos de excitaciónes vocales, las sonoras y las sordas. La excitación sonora se modela mediante la introducción de un tren de impulsos, secuencias unitarias de impulsos separadas por la frecuencia fundamental. La señal sorda se modela mediante un ruido blanco Gaussiano[5] con una media cero y una varianza uno. Un interruptor permite elegir entre sonido sonoro o sordo. A continuación está el filtro todo polos que dará como respuesta la señal de voz.

La función de transferencia del filtro es la siguiente:

$$H(z) = \frac{\sigma}{1 + A(z)} \quad donde \quad A(z) = \sum_{K=1}^{p} a_k z^{-k} \tag{5}$$

los coeficientes a_k (k=1,...,p) son los coeficientes del filtro también llamados LPC (linear prediction coeficients)y σ es la ganancia que nos aporta el filtro. Si lo escribimos de la siguiente manera da lugar a una interpretación más intuitiva:

$$y(n) = \sigma x(n) - \sum_{k=1}^{p} a_k y(n-k) \tag{6}$$

La interpretación que le sigue a esta expresión es que se está usando un predictor lineal cuya función de transferencia es -A(z) y el error de la predicción es $e(n) = \sigma u(n)$.

[5]El ruido blanco Gaussiano es una señal aleatoria cuyos valores no tienen correlación pero siguen la densidad de probabilidad Gaussiana que en este caso es $p_G(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{z^2}{2}}$

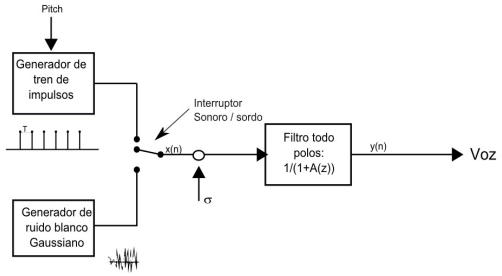

Figura 4. *Modelo LPC de la producción de voz, extraído de [2]*

2.1.3. Procesamiento de la señal de voz

Lo primero que hay que tener en cuenta es que para el procesamiento de la señal de voz estamos usando un medio digital, y como tal no funciona en un entorno continuo como lo es la señal de voz, sino que en un entorno discreto. Por lo tanto lo primero que se hará es convertir la señal analógica en digital. Para ello se muestra la señal con un periodo suficientemente pequeño obteniendo así $x(t) \rightarrow x(nT)$ donde x(t) es la señal en el espacio temporal continuo, n es un número natural y T el periodo con el que se realiza el muestreo. Para facilitar la notación se escribe x(n) en lugar de x(nT) pero hay que recordar que la distancia temporal entre x(n) y x(n+1) será el periodo de muestreo, siendo x(n) la versión discreta de x(t).
Por lo tanto es a la versión discreta x(n) a quien se le aplican el modelo que hemos visto en la sección anterior.

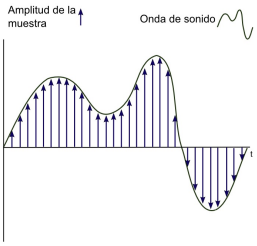

Figura 5. *Versión discreta y continua de la señal de voz.*

Para un procesamiento efectivo de la señal de voz x(n) hay dos requisitos previos que conviene realizar. El primero es la reducción de ruido de la señal y la aplicación de un filtro [2] para enfatizar la señal y aumentar la parte alta de frecuencias, debido a que estas frecuencias tienen un decaimiento de rápido (debido a su paso por el tracto vocal). La audición es más sensible a señales por encima de 1KHz y si no se aplicase el filtro prodríamos tener problemas a la hora de analizarlas[1].

$$P(z) = 1 - \mu z^{-1} \quad siendo \quad \mu < 1 \tag{7}$$

Función de transferencia del filtro para el énfasis de la señal.

El segundo paso es la segmentación de la señal. Tanto la extracción de características como su clasificación es mucho más efectiva cuando se usa una ventana adecuada. Por lo tanto se segmenta la señal que se quiere evaluar en fragmentos de 20-40 ms denominados

frames. Los frames se diseñan de forma que queden solapados y que el desfase entre frames sucesivos sea T_p, el periodo de muestreo.
En ese pequeño intervalo de tiempo la señal se puede considerar constante y por lo tanto obtendremos fracciones cuasi estacionarias. A cada frame se le asignará su vector de características, para que su posterior clasificación resulte más sencilla.

2.2. Modelos de emociones

A la hora de describir un proceso científicamente la clasificación juega un papel fundamental. Por lo tanto si se quiere crear una forma de detectar las emociones en el habla es importante usar un modelo de emociones adecuado.

Existen dos vertientes a la hora de modelar las emociones, la primera es la vertiente clásica usada en los inicios de la detección de emociones, el modelo discreto. La segunda es un modelo más actual y complejo capaz de abarcar un abanico más amplio de emociones, el modelo continuo.

2.2.1. Modelo discreto

Según el modelo discreto existen un número finito de emociones principales y un número más amplio de emociones secundarias que son creadas por combinaciones de las primeras. Este es un modelo útil a la hora de identificar emociones básicas, pero que dificulta la tarea de la identificación de emociones complejas o similares entre ellas.

- Emociones primarias: El enfado, la alegía, la tristeza, el miedo y el disgusto u odio.
- Emociones secundarias creadas a partir de combinaciones de las anteriores, incluyen: La pena, la ternura, la ironía y la sorpresa....

El hecho de clasificar las emociones como básicas y secundarias es una tarea complicada y de difícil consensuación, por lo que muchas veces se abandona este modelo y se fomenta el modelo continuo.

2.2.2. Modelo continuo

Este es un modelo mucho más complejo en los que las emociones se representan en espacios n-dimensionales, dónde los ejes representarán las primitivas emocionales. Estas primitivas pueden variar dependiendo del modelo. Generalmente, sin embargo, se utilizan las siguientes[14]:

- Dominación:

  Mide el origen de las emociones, es decir, si se originan en el sujeto o las genera el ambiente.

- Valencia:

  Nos dará información sobre el grado de positividad o negatividad de la emoción.

- Actividad:

  Mide el grado de intensidad de la emoción.

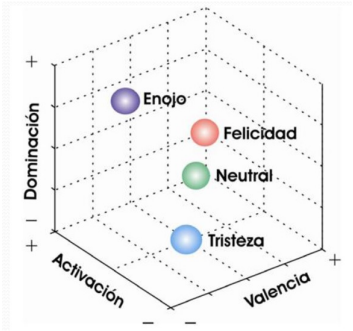

Figura 6. *Mapa de algunas emociones en el modelo continuo, extraído de [16]*

Partiendo de estas tres primitivas se crea un espacio en el que cada emoción tiene su nivel particular de cada primitiva, de esta manera se puede crear un mapa tan complejo como se desee.

Como podemos observar en la Figura 6, tanto la felicidad como el enfado, las dos emociones que analizaremos, tienen un grado de potencia alto es decir que se originan mayormente en el sujeto. Ambas son emociones de gran activación, pero la mayor diferencia entre ambas, como cabe suponer es la valencia, mientras que en el enfado es pequeña, ya que se trata de una emoción negativa, en la felicidad es elevada.

2.3. Características de la señal de voz

En la sección de la voz hemos nombrado algunas de las carácteristicas más importantes de la voz. En esta sección analizaremos las características de la señal de voz que, de acuerdo con la bibliografía, relacionaremos con las emociones. Para ello se han tomado como referencias básicas [14], [15], [16], [5], [12], [13] y [18]. Como se muestra en esta bibliografía las diferentes características se pueden agrupar en los siguientes grupos:

- Prosodia
- Características espectrales
- Calidad de voz

Además se hará un ejercicio de visualización de estas características con datos reales extraídos de las llamadas facilitadas.

2.3.1. Prosodia

La prosodia es la rama de la lingüística que analiza y representa formalmente los siguientes elementos de la expresión oral:

- Entonación:

  La entonación es la variación de la frecuencia fundamental o pitch. En el mecanismo de producción de voz hemos visto que el pitch se puede definir en el marco temporal como el periodo fundamental representado por el lapso de tiempo entre dos pulsos laringeales consecutivos. Para hacer un análisis matemático de la voz sin embargo, es preferible fijarse en la característica cuasi-harmónica de la voz que como hemos

comentado viene por la forma casi triangular de la onda de voz [23]. Por lo tanto definiremos el pitch como la frecuencia representada en el patrón de la voz.
La relación entre frecuencia fundamental y periodo fundamental es $F_0 = 1/T_0$.

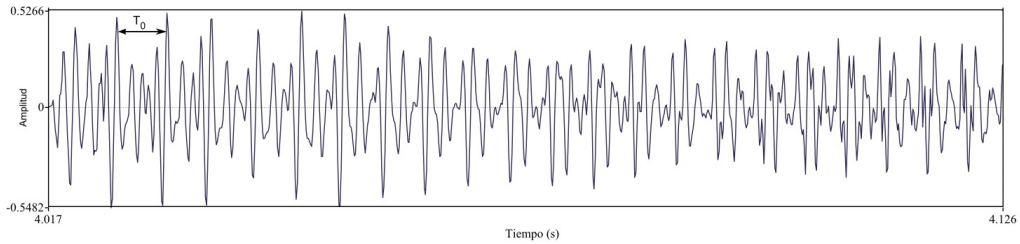

Figura 7. *Forma de la señal de voz al pronunciar la sílaba "*ol*", en la cual se distinguen los pulsos laringeales, extraído de la llamada que expresa alegría.*

Se ha observado que tanto el valor medio como el rango dan información sobre el nivel de activación del orador, correspondiendo una frecuencia fundamental alta a un grado de activación mayor. Así mismo, las fluctuaciones nos darán información de la positividad de la emoción, así una emoción positiva tendrá una fluctuación del pitch suave, mientras que en una negativa variará de una forma más brusca.

A la hora de estudiar el pitch hay que tener en cuenta la diferencia entre la voz masculina y femenina, tanto la media como la desviación estándar suelen ser mayores en la voz femenina.

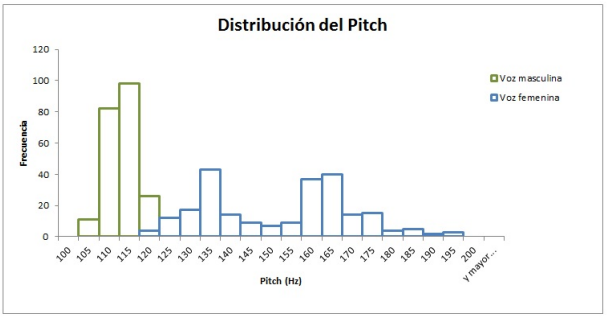

Figura 8. *Ejemplo de la distribución del pitch perteneciente a la voz femenina donde en este caso la media es 150 Hz y el perteneciente a la voz masculina donde la media en este caso es 110 Hz, extraídos de la llamada de alegría y la segunda llamada de enfado respectivamente*

- Articulación:

  La articulación está directamente relacionada con los formantes y su posición. Los formantes son las resonancias del tracto vocal y nasal que aparecen frecuentemente en espectrogramas de regiones de alta energía y que varían lentamente en el tiempo. Estas dos características son conocidas en las vocales. Por eso los formantes se entienden como picos espectrales representados en el campo de frecuencias de la voz que corresponden a subyacentes resonancias vocales.
  Los formantes son muy útiles, puesto que pueden describir fácilmente la forma del

tracto vocal. Al igual que el pitch son una característica intrínseca, pero pueden variar su posición dependiendo de la emoción del orador.

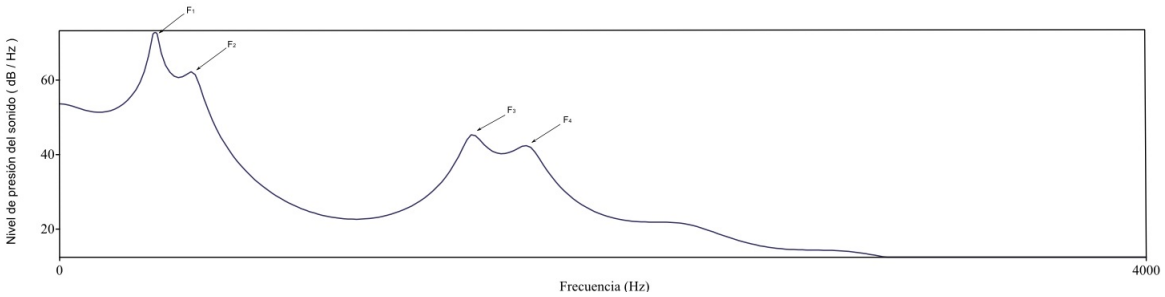

Figura 9. *Formantes de la palabra "*dije*", extraído de la primera llamada de enfado*

En las Figuras 9 y 10 se muestran dos formas de representar los formantes gráficamente. En la Figura 10 la intensidad del color muestra la intensidad de la voz. Si no hubiese ningún ruido en el espectrograma sólo se marcarían las posiciones de los formantes. Por otra parte y uniéndolo con el apartado anterior, señalamos que las finas líneas verticales que se observan son los pulsos glotales que dan la medida del pitch.

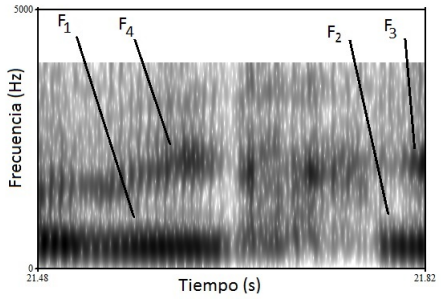

Figura 10. *Espectrograma del mismo fragmento que la figura anterior*

- Intensidad:

  La intensidad mide el volumen de la voz, es decir cuan alta o baja es la voz. Hemos visto que la intensidad está condicionada por la presión subglotal puesto que ésta determina la velocidad del flujo de aire que atraviesa la tráquea.
  Hay dos factores en la vibración de las cuerdas vocales que determinan la intensidad. El primero es la cantidad de vibraciónes de las cuerdas vocales: A menor número de vibraciones menor intensidad tendrá la voz. El segundo la amplitud de las vibraciones: Cuanto mayor sea la amplitud con la que vibran las cuerdas vocales mayor será la intensidad de la voz, puesto que para que vibren con una mayor amplitud se necesita un mayor flujo de aire o lo que es lo mismo una mayor presión subglotal. Matemáticamente la intensidad se representa por la energía de la señal ya que $Intensidad = \frac{Energía}{Tiempo * Área}$, el área se calcula a partir de la distancia entre la fuente de energía y el lugar desde dónde se mide.
  El oído humano mide el sonido en una escala logarítmica, por lo tanto es común escribir

la energía de acuerdo con esta escala.

$$E = 10 \log \left(1/N \sum_{n=0}^{N-1} x^2 [n] \right) \tag{8}$$

donde x[n]es como hemos dicho anteriormente la versión discreta de x(t) y N es el número total de muestras.

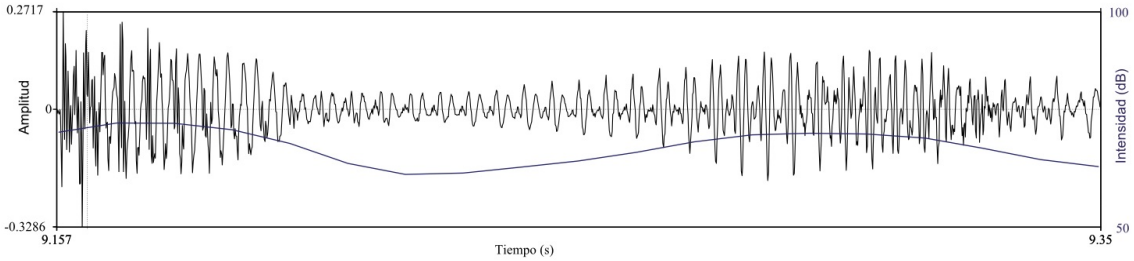

Figura 11. *Señal de voz e intensidad de la palabra "*tengo*", extraída de la llamada con alegría*

En la Figura 11 se observa la estrecha relación entre la onda de voz y la intensidad de la misma. Así mismo se ve con claridad como baja la intensidad entre las dos sílabas que componen la palabra.

Se ha observado que la intensidad de la señal está directamente relacionada con la activación, así a mayor intensidad mayor será la activación del orador.

- Pausas:

  La cantidad de pausas realizadas en una frase se puede medir fácilmente creando un algoritmo que discrimine la presencia de la ausencia de voz. Generalmente se usan dos tipos de algoritmos para esta tarea [1].
  El primero se basa en la medición de la presencia de pitch, esta es una condición suficiente para la presencia de voz pero no necesaria, puesto que puede que exista voz en un fragmento aunque el pitch no sea medible.
  Otra forma más fiable es medir el ratio de energía armónica respecto a la no armónica, esta última creada por el ruido que pudiera haber. Un umbral ampliamente usado es el siguiente:

$$ratio > 10dB \rightarrow Voz \tag{9}$$

$$ratio < 4dB \rightarrow Silencio \tag{10}$$

  En la Figura 12 se expone un fragmento del segmento en el que el orador dice la frase *Obviamente no*, en la cual se muestra el pitch en azul y la intensidad en verde. Se observa como en el fragmento marcado con lineas discontinuas el pitch es nulo. Este tramo correspondería a un silencio si se usase el primer algoritmo, sin embargo si vemos la forma de la onda de voz vemos que no puede corresponder a un silencio y de hecho al medir la intensidad vemos que la media en este segmento es 76.1 dB. Al medir el ratio de energía armónica con respecto a la no armónica vemos que se sitúa en el rango de voz.

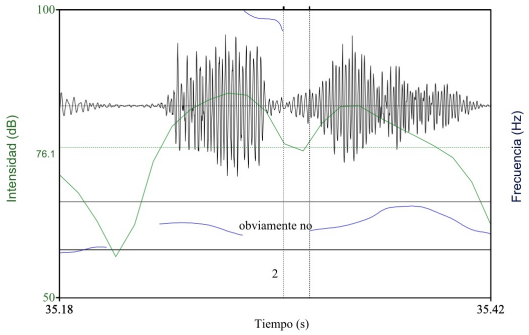

Figura 12. *Ejemplo de pitch nulo pero que no corresponde a un silencio, fragmento de "*obviamente no*",extraído de la primera llamada con enfado.*

El número de pausas dan información sobre la positividad de la emoción. Se ha observado que en un discurso triste el número de pausas es mayor que en un discurso alegre. La duración de las pausas también tiende a ser mayor en un discurso de menor valencia.

- Ritmo:

  El ritmo o velocidad mide la rapidez con la que habla el orador. Para poder medir la velocidad en este trabajo se ha realizado la transcripción del discurso. A partir de la transcripción se ha realizado una transcripción fonética.[6] La velocidad de locución se ha estimado como el número medio de fonemas por unidad de tiempo.

  El ritmo da información sobre la activación de la emoción, un ritmo más alto corresponderá a una activación mayor.

2.3.2. Características espectrales

Para obtener un análisis completo de la señal de voz es preciso analizar las características en el dominio de la frecuencia, para ello habrá que trasladar la señal desde el dominio temporal a este dominio, la forma más común de de hacer esto en cualquier tipo de señal o función que reside en el dominio temporal es haciendo una trasformada de Fourier. En este caso como nuestra señal es una señal discreta tenemos que usar una transformada de tiempo discreta de Fourier:

$$X(e^{i\omega}) = \sum_{n=-\infty}^{\infty} x\,[n]\, e^{i\omega n} \tag{11}$$

donde x[n] es la equivalente discreta de la señal temporal, ω la frecuencia y X($e^{i\omega}$) la transformada de la señal discreta en el dominio de las frecuencias.

La transformación de Fourier es el método más ampliamente utilizado para trasladar las funciones de un dominio a otro. En este caso existen sin embargo otro tipo de tratamientos que nos darán coeficientes más apropiados para el análisis de señales de voz [7] como lo son los coeficientes cepstrales o los coeficientes de predinción lineal. Estos dos coeficientes vienen

[6]El fonema es la unidad lingüística básica, los fonemas de un idioma construyen un número finito de sonidos distinguibles y mutuamente excluyentes

del análisis cepstral que veremos a continuación.

Análisis cepstral:
Los cepstrales vienen definidos de la siguiente manera:

$$c\,[n] = \mathscr{F}^{-1}\{log|\mathscr{F}\{x[n]\}|\} \tag{12}$$

Es decir, la transformada inversa de Fourier, denotada por $\mathscr{F}^{-}$, del logaritmo absoluto del espectro de la señal de voz, es decir de la transformada discreta de Fourier, denotada por $\mathscr{F}$, de la señal x[n]. Hay que resaltar que los coeficientes cepstrales vuelven a estar en el dominio temporal, o más precisamente pseudotemporal.

En el caso de un frame de una señal de voz lo que hay que hacer es:

$$c[n] = \sum_{n=0}^{N-1} log\left(|\sum_{n=0}^{N-1} x[n]e^{-i\frac{2\pi k}{N}n}|\right)e^{i\frac{2\pi k}{N}n} \tag{13}$$

donde k representa el número de onda.
Se trata de un operado que transforma una comvolución en el dominio temporal entre la excitación y el tracto vocal en una suma en el dominio cepstral.

Hay dos razones por las cuales se usa este procedimiento, primero por que al calcular el logaritmo del espectro se reduce tanto el rango dinámico como la diferencia de amplitud entre los armónicos , y segundo porque al realizar la trasformada discreta de Fourier y suponer que el logaritmo de la misma también tiene forma de onda esperaremos que esta sea cuasi-periódica con modulaciones de su amplitud.

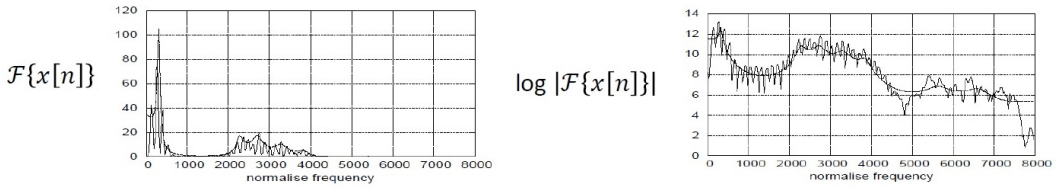

Figura 13. *Variación de la frecuencia, extraído de [7].*

Al trasladarlo al dominio temporal esto se mostrará mediante un pico cerca del periodo de la señal y componentes a bajas frecuencias relacionados con la modulación. En este dominio podemos separar estas dos atribuciones. Si queremos fijarnos en la excitación glotal mantendremos los coeficientes cepstrales altos y si por el contrario nos queremos centrar en las características del tracto vocal mantendremos los coeficientes cepstrales bajos. Por su cualidad de enfatizar la periodicidad de los armónicos, los coeficientes cepstrales son muy útiles a la hora de detectar la presencia de pitch.

En la figura 14 se observan perfectamente el pico relacionado con la excitación glotal y los componentes de bajas frecuencias relacionados con la modulación.

- Coeficientes cepstrales de la escala de Mel:

  La obtención de estos coeficientes es muy similar al proceso de obtención de los coeficientes cepstrales con la singularidad de que estos coeficientes son capaces de representar el habla basándose en la percepción auditiva humana. Para lo cual hay que

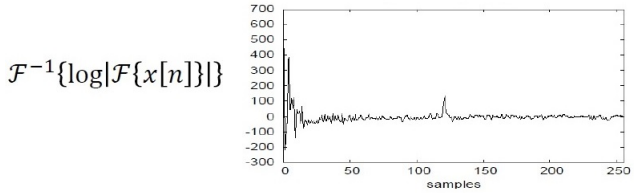

Figura 14. *Forma de los coeficientes cepstrales, extraído de [7].*

incorporar un filtro con escala logarítmica, escala a la que percibe los sonidos el humano, denominado escala de Mel.

$$M(f) = 1125ln(1 + f/700) \tag{14}$$

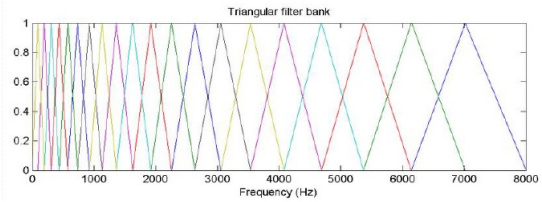

Figura 15. *Filtro triangular Mel, extraído de [21].*

Para obtener estos coeficientes primero aplica el logaritmo a la transformada discreta de Fourier, al igual que en el caso anterior, posteriormente se aplica el filtro Mel y por último se calcula la transformada de coseno discreta.

$$MFCC[n] = 1/N \sum_{n=0}^{N-1} x'[k] \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right] \tag{15}$$

donde k es el número de onda, x'[k] es el resultado de aplicar la transformada discreta de Fourier y el filtro Mel a la señal de voz discreta x[n], N la longitud del frame y MFCC[n] los coeficientes cepstrales de mel que representan la amplitud del espectro de la señal en base a la percepción humana.

- Coeficientes cepstrales de predicción lineal:

  En secciones anteriores hemos hablado sobre los coeficientes de predicción lineal que son capaces de estimar el envolvente espectral de la señal con gran precisión, sin embargo estos coeficientes son muy sensibles a la precisión númerica, por ello es conveniente transformarlos a coeficientes cepstrales de predicción lineal. El cálculo de estos coeficientes se hace de manera recursiva con el siguiente algoritmo:

$$LPCC[n] = \begin{cases} ln(P) & n = 0 \\ a[n] + \frac{1}{n}\sum_{k=1}^{n-1} kLPCC[k]a[n-k] & 1 < n \end{cases} \tag{16}$$

donde a[n] son los coeficientes lineales obtenidos, LPCC[n] los coeficientes cepstrales de predicción lineal y P un número por determinar a la hora de calcularlo dependiendo del error que se precise [11].

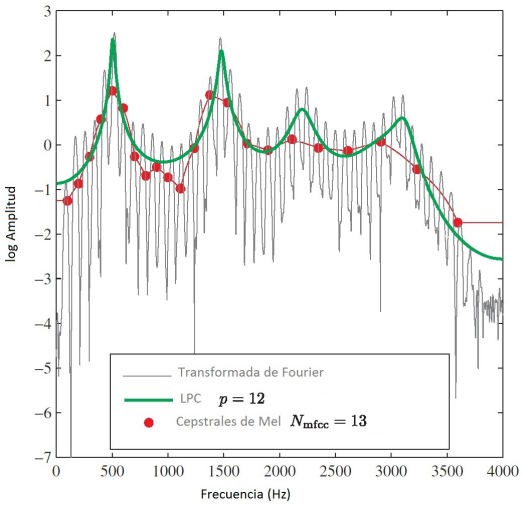

Figura 16. *Comparativa de los coeficientes cepstrales de mel y los de predicción linear[7]*

2.3.3. Calidad de voz

La calidad de voz es la característica que nos permite diferenciar las voces de las distintas personas, es la *imagen* de la voz.
La calidad de voz está formada por un conjunto de característica, normalmente se incluyen la velocidad, la intensidad y el pitch. Esta característica está condicionada por las diferencias anatómicas del individuo, que explican las diferencias de voz entre distintos individuos, como es el claro ejemplo de la diferencia de voz entre el sexo femenino y el masculino, y por aspectos extralingüísticos como la forma de hablar habitual de esa persona o la emoción que está expresando, caso que es de nuestro interés.

La calidad de voz es difícil de cuantificar y calificar, pero en general se suele hablar de los siguientes cuatro grupos [8], [9], [6] y [24]:

- Voz entrecortada y voz susurrante:

  En estos dos tipos de voces se escucha una fricción característica, puesto que el aire atraviesa audiblemente las cuerdas vocales y se exhala casi en su totalidad por la boca. La diferencia entre estos dos tipos de voces reside en la tensión de los músculos respiratorios. En la voz entrecortada los músculos están relajados, permitiendo que las cuerdas vocales no se cierren durante la vibración, por lo que el flujo de aire entre las cuerdas vocales es grande, creando la fricción audible. Por el contrario en la susurrante los músculos respiratorios están tensos y, aunque la porción media de las cuerdas vocales esté cerrada, se mantiene una abertura triangular que permite que el flujo de aire se expulse. Esta abertura es inversamente proporcional a la fricción que se escucha.

  Detectar estos dos tipos de voz es complicado y sobre todo es complicado diferenciarlos entre ellos. Una característica resaltada en distintos estudios es la presencia de mayor ruido anarmónico, especialmente en el área del tercer formante, y una atenuación de las altas frecuencias.

  Así mismo también es complicado marcar las emociones que determinan estos dos

tipos de voz. Se ha observado que ambos pertenecen a tipos de emociones de baja activación y especialmente la voz susurrante puede denotar inhibición en el orador.

- Voz rota *(Vocal fry)*:

  La voz rota se caracteriza por una frecuencia fundamental muy baja y la percepción de la vibración de las cuerdas vocales.
  Se produce al dejar suelto el cierre glotal, lo que permite que el aire atraviese la traquea muy lentamente haciendo sonar un traqueteo a una frecuencia muy baja. El cartílago aritenoides, cartílago en el que se insertan las cuerdas vocales, y la laringe se juntan haciendo que las cuerdas vocales permanezcan comprimidas fuertemente convirtiéndolas en una masa floja y compacta. Esto hace que vibren pronunciadamente y de forma irregular creando el sonido característico de este tipo de voz cuando el aire atraviesa el cierre glotal.

  Una característica muy apreciable es la medición de un pitch excesivamente bajo y de periodos irregulares. Así mismo el pulso glotal también disminuye hecho que hace que el espectro sea relativamente plano.

  Parece ser que el uso de la voz rota indica aburrimiento y/o indiferencia por parte del orador.

- Voz áspera:

  La voz áspera es el resultado de una gran tensión en las cuerdas vocales que tiene como consecuencia la aproximación excesiva de las cuerdas vocales. Esta tensión hace que la laringe quede encogida haciendo que las cuerdas ventriculares estén en contacto con la superficie alta de las cuerdas vocales imposibilitando la vibración.

  Cuando hablamos del mecanismo de la producción de voz se dijo que la forma de onda tiende, con un alto grado de aproximación a una forma triangular. En la voz áspera esta forma se vuelve más irregular debido al proceso arriba mencionado. Esto se traduce en aperiodicidades tanto en las frecuencias como en la amplitud de la señal acústica. Las aperiodicidades son fluctuaciones de corto plazo que ocurren de un periodo a otro, las aperiodicidades que ocurren en la frecuencia se les llama perturbaciones de frecuencia o jitter y a las que ocurren en la amplitud perturbaciones de amplitud o shimmer.
  Estos dos fenómenos junto con el aumento de ruido interarmónico en bajas frecuencias son muy útiles a la hora de detectar este tipo de voz.

  El tipo de emociones que desencadenan la voz áspera son emociones de baja valencia y alta activación.

- Voz modal:

  Es el registro vocal más usado, en el cual la combinación óptima de flujo de aire y tensión glotal produce una vibración máxima y cuyo modelo de producción estudiamos en anteriores secciones.

2.4. Selección de características y procesado de características

La selección de las características adecuadas es primordial a la hora de crear un sistema eficaz computacionalmente. En general existen dos formas de selección:

- Selección manual: Es posible, basándonos en trabajos previamente realizados observar qué características son las que están más directamente relacionadas con las emociones que vamos a analizar y centrarnos en estas. Este método es razonable cuando se

conocen muy bien las emociones a detectar ya que el coste computacional decrece considerablemente, sin embargo, en la práctica no es tan sencillo, puesto que no existe un acuerdo de cómo determina la variación de una característica en la clasificación de una emoción. En cada caso particular existe un determinado grupo de caracerísticas distinto que parece que tiene más correlación con la emoción.

- selección automática: Para la selección se usa un algoritmo que busca el grupo características que más correlación tiene con la emococión que estamos buscando y menos correlación entre ellas. Así evitaremos la redundancia que acarrea tanto un coste computacional innecesario como una mayor probabilidad de error a la hora de la clasificación.
  La ventaja de esta selección es que no hay que hacer ninguna selección previa de las características que vamos a evaluar, por lo tanto no perdemos información.

Para concluir con este apartado veremos las dos formas que hay de procesar las características que hemos visto, la primera es localmente, es decir, analizar el vector de características de cada frame individualmente. Tras la clasificación se le asignará una puntuación que determinará la emoción del frame.
La segunda forma es procesarlas globalmente, en este caso se tomarán medidas estadísticas que se determinen adecuadas de cada característica entre todos los frames de un enunciado. Estas medidas estadísticas incluyen la media, la desviación estándar, los cuartiles, los máximos y mínimos y cualquier otra que por las peculiaridades de cada característica de la voz se considere oportuna. Con estos datos se creará un nuevo vector de características que en este caso en vez de ser por frame será por enunciado, de la misma manera que en el caso anterior este vector se clasificará, se le asignará una puntuación y en base a esta puntuación se determinará la emoción presente en la frase.

La diferencia entre estas dos formas de procesamiento es obvia, mientras que en la primera la evaluación de las emociones se hace por fragmentos de entorno a los 20 ms de duración en la segunda se hace por fragmentos de varios segundos de duración, por lo tanto cualquier variación de emociones en una misma frase no se podrá detectar mediante un análisis global pero si mediante uno local.

En lo que a los resultados obtenidos por cada uno de los dos métodos de procesamiento, la mayoría de investigadores [15] coinciden en el hecho de que es más beneficioso procesar las características globalmente. Tanto por su precisión en la clasificación como por el tiempo de ejecución de los clasificadores que cada uno requiere, debido a que el número de vectores de características es mucho menor, puesto que tenemos uno solo por frase, el tiempo que necesita el algoritmo para la clasificación es menor.
Sin embargo, aunque según los estudios realizados se ha demostrado la fiabilidad de las características globales para diferenciar estados de ánimo con diferente activación como la alegría y la tristeza, no existen pruebas concisas de que también sean útiles a la hora de clasificar emociones con un nivel de activación similar, como la alegría y el enfado. Por otra parte, como hemos mencionado antes al tratarse de medidas estadísticas dentro de una frase, se pierde toda la información temporal que nos daría un análisis de características locales.
Por último, aunque la ejecución de los algoritmos en el procedimiento global sea más rápido el precio que se tiene que pagar es que el número de vectores de entrenamiento usados por el algoritmo también es menor y esto puede impedir el uso de determinados clasificadores complejos.

2.5. Test de significación estadística

Antes de concluir la sección teórica y dar paso a la parte de la memoria dónde se analizaran en profundidad los datos facilitados, queremos aprovechar a introducir varios conceptos estádisticos que se usarán posteriormente en el análisis de las señales.

Realizaremos tres test estádisticos a los fragmentos de cada señal que muestren más claramente el estado neutro y el estado de la emoción al que pertenecen. El objetivo de estos test no es otro que saber si la distribución estadística de los parametros escogidos en segmentos neutros y emocionales es significativamente diferente.

- El primero de los test que se realizará es el test Z bilateral con una muestra. El objetivo de este test es ver la confianza de la medida, es decir, saber cual es el intervalo que incluye la media con un 95 % de probabilidad. Este intervalo será el intervalo de confianza de la muestra, en este caso al ser la probabilidad de éxito $((1-\alpha)100)$ el 95 % $\alpha = 0,05$, al valor α se le denomina nivel de significación o error aleatorio.
  Primero tenemos que saber cual es la distribución teórica del parámetro cuyo intervalo de confianza queremos conocer. Es común suponer una distribución normal,[7] puesto que con una población suficientemente grande la distribución de las muestras es prácticamente una distribución normal con una media μ y una desviación típica dada por $\frac{\sigma}{\sqrt{n}}$ denotada por $\bar{X} \sim N(\mu, \sigma)$. Por lo que:

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = Z \sim N(0,1) \tag{17}$$

  donde $\bar{X}$ es la distribución de nuestra muestra, μ, σ y n son la media, la desviación típica y el tamaño de la misma respectivamente, Z es la nueva distribución y N(0, 1) la distribución normal o gaussiana de media 0 y desviación típica 1.

  En esta nueva distribución es sencillo calcular los límites superiores e inferiores para que la media se situe dentro de este intervalo con una probabilidad de $(1-\alpha)100$, estos límites están dados por la expresión:

$$Z_{\pm} = \mu \pm z\frac{\sigma}{\sqrt{n}} \tag{18}$$

  donde μ, σ y n son la media, la desviación típica y el tamaño de la muestra respectivamente y el valor de z está tabulado, en este caso para un intervalo de confianza del 95 % es 1,96.

  Tras efectuar este test se puede inferir en la distribución de dos muestras distintas, así si los intervalos de confianza de dos muestras se solapan se deduce que existe la probabilidad de que las medias de las dos muestras residan en el mismo intervalo por lo que la diferencia entre la distribución entre ambas no es estadísticamente significativa.

[7] Una distribución normal es aquella en la que la media es cero y la desviación típica 1

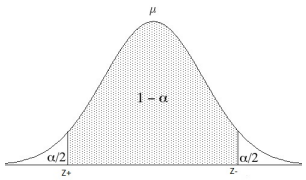

Figura 17. *Muestra la densidad de probabilidad* $1-\alpha$ *de una distribución cuya media es* μ *entre los puntos* $Z_{\pm}$*, fuente de la figura wikipedia*

- El segundo test es el test Z bilateral con dos muestras, este tiene como objetivo decidir si la diferencia de los parámetros en los segmentos neutros y en los segmentos que muestran una emoción (alegría o enfado) es estadísticamente significativa o no. Para ello se calculará el siguiente intervalo:

$$Z_{\pm} = (\mu_1 - \mu_2) \pm z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{19}$$

  donde 1 y 2 denominan la primera y la segunda distribución y μ, σ y n la media, la desviación típica y el tamaño respectivamente.

  Si dentro de este intervalo se encuentra el 0, la diferencia entre ambas distribuciones no es estadísticamente significativa. Para que el intervalo no contenga el cero la diferencia de las medias tiene que ser grande con respecto a la distribución típica normalizada. Este test se realizará tanto por segmentos como por emociones enteras.

- En los casos donde la diferencia de ambas distribuciones no sea estádisticamente significativa, se procederá a hacer el tercer test. Este test tiene como objetivo saber si se puede rechazar o no la hípotesis nula, la hipótesis nula H_0 la cual no postula diferencias estadísticamente significativas entre las medias y si las hay estas se deben al azar, la hipótesis alternativa H_1 se contrapone a esta hipótesis, por lo que aceptar una rechaza automáticamente la otra. El error al aceptar la hipótesis nula será 0,05.
  Para completar este test se calculará el valor F de la muestra que viene dado por la razón de la varianza de las dos muestras que queremos estudiar.

$$F_{prueba} = \frac{\sigma_1^2}{\sigma_2^2} \tag{20}$$

  siendo σ^2 las varianza de la primera y segunda muestra. Seguidamente se compara el valor obtenido con el valor tabulado de F para los grados de libertad del numerador, denominador y valor de α que tenemos. Si $F_{prueba} < F_{tabla}$ se acepta la hipótesis nula y sino se rechaza. En nuestro caso $\alpha = 0,5$, los grados de libertad del numerador es $k-1$ siendo k el número de muestras, por lo que en nuestro caso será 1 y los grados de libertad del denominador $k(n-1)$ siendo n la longitud de nuestra, por lo que en este caso variará dependiendo del caso que tengamos.

3. Procesamiento de la señal y extracción de características

Terminada la sección teórica, nos volcaremos en el análisis de las tres llamadas telefónicas facilitadas por la empresa Xupera.

En la actualidad existen distintas herramientas [20] y [22] capaces del procesamiento de voz y extración de sus características. En este trabajo nos hemos decantado por el programa libre *Praat* [22],creado en la Universidad de Amsterdam por los expertos en ciencias fonéticas David Weenink y Paul Boersma. Este programa permite etiquetar las señales de voz, extraer las diferentes características personalizando el análisis y graficar los resultados entre muchas otras cosas.

Antes de comenzar el procesamiento de las señales de voz facilitadas, se hizo una primera escucha para valorar su calidad. Tras esto se hizo un procesamiento manual de cada una de las señales. Se comenzó con un primer etiquetado de los silencios en la señal. Para ello se utilizó la herramienta del programa que incorpora el algoritmo 9 y 10 de la sección anterior. En cada caso se tuvo que ajustar el rango para el cual se etiquetaría como silencio dependiendo del ruido existente en cada señal. Así para las señales donde había más ruido el rango era mayor. La extracción de características en este programa se tiene que hacer manualmente, seleccionando uno a uno los segmentos de los cuales se quieren extraer las características y seleccionando cada una de las que quieres extraer. Para facilitar esta tarea se optó por añadir al etiquetado de los silencios las etiquetas de operador, ruido y música para dejarlas fuera del análisis. Con el mismo objetivo se creó la etiqueta de ambos, la cual delimitaba los tiempos en los que las dos voces estaban presentes, por lo que en el análisis las características de la voz del cliente se verían mezcladas con las del operador. Hecho ésto se habían conseguido separar los fragmentos de voz válidos para el análisis. En este punto se decidió hacer una transcripción de cada uno de los fragmentos útiles, acción que sería útil tanto para poder diferenciar los segmentos más fácilmente, como para medir la velocidad de los mismos tal y como se explicó en seccion 2.3.1.

Por último se hicieron dos etiquetados independientes de las emociones de cada fragmento, uno hecho por mí y el otro por mi compañero. Este etiquetado procedió en dos partes, primero se hizo una discriminación entre alegría y neutro o enfado y neutro, dependiendo de la llamada. Después dentro de la emoción a detectar en esa llamada, a saber alegría o enfado, se valoró el segmento que expresaba esa emoción con una puntuación del 1 al 3, siendo 1 una menor manifestación de la emoción y 3 una manifestación muy notoria. Esta segunda división es útil tanto para hacer un clasificador más preciso, como para poder seleccinar en el análisis que se realizará a continuación sólo aquellos fragmentos donde la emoción sea más notoria.
La detección de emociones en el habla es una tarea complicada incluso para los humanos y aunque en fragmentos donde la emoción está muy marcada es fácil de distinguir y entre los dos etiquetados independientes había una concordancia del 100 %. La cosa cambia cuando la emoción no es tan evidente, por lo tanto para una mayor fiabilidad sólo se escogieron para el siguiente proceso aquellos fragmentos donde los dos etiquetados coincidían. Esta coincidencia podía ser total o parcial. La total ocurre cuando coinciden tanto la etiquetación de emoción como el nivel de la misma y la parcial cuando sólo coinciden el tipo de emoción. Los fragmentos con coincidencia parcial se usarán sólo para entrenar al algoritmo en la discriminación entre neutro y emoción.

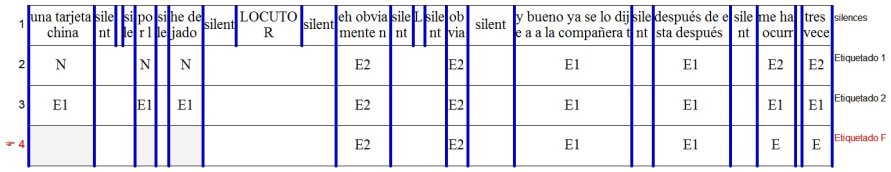

Figura 18. *Ejemplo de los dos etiquetados y el etiquetado final*

En la Figura 18 se ve un ejemplo de como fue la selección del etiquetado final, en la cual se aprecian segmentos con coincidencia total, parcial y ninguna coincidencia. En la Figura 19 el resultado final de otra de las señales.

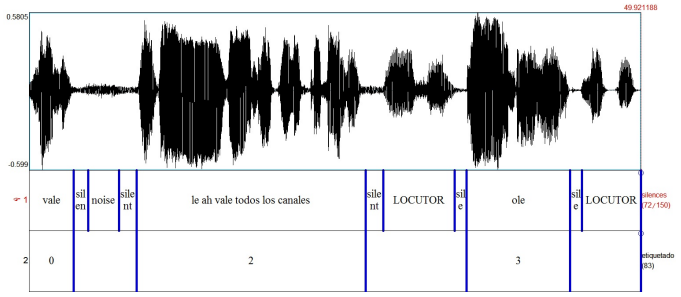

Figura 19. *Ejemplo del etiquetado*

Hemos visto que hay muchas características presentes en la señal de voz. Sin embargo no podemos extraer todas ellas puesto que es muy complejo usar un algoritmo que tenga en cuenta todas y el coste computacional también sería muy elevado. Por lo tanto hemos optado por elegir las más relevantes y centrar nuestro análisis en estas. Hay que tener en cuenta que el objetivo no es poder diferenciar alegría de ira, sino alegría de neutro y ira de neutro. Por un lado no tenemos ninguna grabación de un mismo orador donde las dos emociones estén presentes. Por otro hemos visto que hay características que en la expresión de ira y alegría se parecen, por lo tanto necesitaríamos una base de datos muchísimo mayor para poder entrenar un algoritmo para diferenciar estas dos emociones. Basándonos en la revisión bibliográfica realizada [14],[15],[16], [18] y [17] decidimos centrar nuestro análisis en la intensidad, pitch, formantes, pausas y velocidad. Primero porque estas son las características más generales y su extracción es más sencilla y segundo porque tanto la alegría como el enfado son emociones con una alta activación y hemos visto como el grado de activación esta muy ligado a estas características.

Las características tienen que extraerse en pequeños frames de 20-40ms donde sean cuasi-estacionarias. El programa da la opción de hacer la división de nuestros segmentos en pequeños frames de varias maneras. La primera es de una manera automática donde según la longitud de cada segmento el programa elige la longitud optima. La segunda es definiendo manualmente la longitud de los frames y la tercera es definiendo la cantidad de frames que queremos que tenga el segmento independientemente de la longitud del mismo. Nosotros elegimos esta última forma puesto que para que el método de clasificación funcione de manera correcta se necesita representar cada segmento por matrices $(n \times m)$ de iguales

dimensiones, una de las dimensiones corresponderá al número de frames y la otra a la longitud del vector de características.

Esta sección tiene como objetivo explicar brevemente los clasificadores usados en el proyecto de matemáticas y que es lo que se buscaba con cada uno de ellos para así poder entender mejor los resultados que se obtengan.

4. Análisis de las señales de voz

Ahora nos centraremos en la evolución de las señales de voz cuando expresan distintas emociones, nos fijaremos especialmente en el pitch, la intensidad, la velocidad y las pausas dentro de una oración. Puesto que estos son los parámetros cuyo variación se puede analizar más fácilmente. Cada sujeto tiene una voz particular, el pitch los formantes y la calidad de voz varían de sujeto en sujeto, por eso en nuestro análisis no mezclaremos las señales, sino que estudiaremos cada señal por separado. Las conclusiones que obtengamos en cada una por lo tanto, no tienen porque poder extrapolarse a las demás señales. Para poder hacer un estudio sobre la detección de emociones sin relación con el locutor se necesitaría en primer lugar una base de datos mayor y en segundo trabajar con las diferencias, así como normalizar respecto a las características de cada locutor en vez de con valores absolutos.

4.1. Grabación con alegría

La llamada que la empresa nos proporcionó donde se aprecia la emoción de alegría es una conversación entre una clienta y un asistente de servicio técnico de Euskaltel. La cliente llama para saber cómo se puede pasar del modo radio al modo televisor desde su mando a distancia.

La llamada tiene una duración de aproximadamente 2.5 minutos y se desarrolla en su mayoría en un tono neutro a excepción de las ocasiones de carácter puntual en las que el sujeto muestra la emoción de nuestro estudio. Estas ocasiones coinciden, como cabe esperar, con el momento en el que consigue solucionar el problema.

Al tratarse de una conversación que se desarrolla por lo general en un tono neutro tanto el pitch como la intensidad permanecen más o menos constantes durante toda la conversación a excepción de las afirmaciones y las interrogaciones en los que el pitch baja y sube respectivamente. Además en el final de cada palabra la intensidad baja hasta valores cercanos a cero en muchas ocasiones. Son comportamientos comunes en el habla humana que se pueden extrapolar a otros individuos e idiomas.

En lo que a la velocidad se refiere, se observa que por lo general la conversación fluye con una velocidad media-rápida (17-22 fonemas por segundo) y sin demasiadas pausas. Aparecen pequeños silencios entre segmentos que contienen de 6 a 12 palabras.

En cuanto a los segmentos etiquetados como alegres hay que tener en cuenta que en esta llamada representan la excepción y tienen corta duración, de una a tres palabras.

Se observa, tal y como se predice en la teoría, que tanto la intensidad como el pitch aumentan en estos segmentos como se observa en la Figura 20. El segmento de alegríaestá representado en azul y su valor medio lo muestran las lineas discontinuas azules. El segmento neutro está representado en verde y su valor medio lo muestran las lineas discontinuas verdes. Ambos segmentos se han elegido de forma que tengan una sola palabra y una duración similar.

En la intensidad del segmento de alegría se muestra una característica llamativa que se repite en otros segmentos con esta misma emoción. Se trata de la prolongación y separación de las sílabas de forma en que se pueda distinguir muy fácilmente cuando termina una y

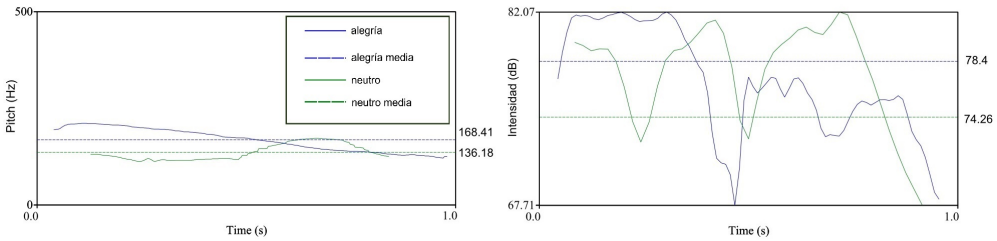

Figura 20. *Distribución de pitch e intensidad en fragmento de alegría "*ole*", y neutro "*mira*"*

empieza la siguiente. Este rasgo también se ve trasladado a los valores del pitch donde en la primera mitad tiene un valor significativamente superior al de la segunda. Esta prolongación de las sílabas se traduce en una velocidad mucho menor en los tramos de alegría (3-8 fonemas por segundo).
El estudio de la cantidad de pausas no es factible en este caso en particular, puesto que como hemos mencionado los segmentos que muestran alegría son pequeños fragmentos y no hay ejemplos suficientes para saber si los silencios están creados por la representación de la emoción o por la misma conversación ya que la presencia del operador es notable en estos casos.

Cabe destacar que el análisis de estas características (intensidad, pitch y velocidad) por separado no es concluyente, ya que se encuentran fragmentos como es el caso de las muletillas ("es que"por ejemplo) que tienen una velocidad similar a la de los segmentos de alegría, u otros segmentos que por la semántica de la frase o presencia de emociones mucho más sutiles pueden tener un pitch o una intensidad más elevada que el resto, como es el caso de las frases interrogativas, pero que no deben de ser clasificados como una expresión de alegría. Por lo tanto es importante fijarse en los valores de un conjunto de características en vez de la evolución de una sola.

Ahora procederemos a analizar los resultados obtenidos tras realizar los test estádisticos a los distintos segmentos seleccionados.

	Pitch (Hz)	
	Alegría	Neutro
μ	220,49	142,26
σ	62,05	40,46
1º test $[Z_+, Z_-]$	[224,47 , 216,52]	[145,83 , 138,69]
2º test $[Z_+, Z_-]$	[83,54 , 72,92]	

	Intensidad (dB)	
	Alegría	Neutro
μ	75,54	68,35
σ	6,54	8,65
1º test $[Z_+, Z_-]$	[76,22 , 74,85]	[69,16 , 67,54]
2º test $[Z_+, Z_-]$	[8,24 , 6,13]	

Tabla 1. *Resultados de los test al comparar la alegría con el estado neutro*

La Tabla 1, parece que tanto la desviación típica de la distribución como la media del pitch cuando se expresa alegría son mayores. Los valores de la intensidad también parecen ser mayores pero con una distribución más pequeña. En cuanto a los resultados del primer test, vemos que los intervalos no se solapan en ninguno de los dos casos, por lo que tenemos una probabilidad mayor del 95 % de que las medias se encuentren en intervalos distintos. Si reparamos a los intervalos obtenidos por el segundo test vemos que no contienen el cero, por lo que la diferencia de las medias es estadísticamente significativa.

Haciendo el primer test para cada segmento obtenemos lo siguiente:

1º test [Z_+, Z.]	1º seg. A	2º seg. A	3º seg. A	1º seg. N	2º seg. N
Pitch (Hz)	[240,24 , 227,00]	[268,07 , 256,89]	[171,03 , 164,55]	[141,68 , 129,04]	[152,30 , 147,74]
Intensidad (dB)	[75,74 , 73,66]	[78,84 , 76,66]	[76,94 , 74,76]	[68,55 , 66,42]	[70,26 , 67,89]

Tabla 2. *Resultados del primer test en los distintos segmentos de alegría y neutro*

En la Tabla 2 observamos que tampoco en este caso el intervalo de confianza de los segmentos con alegría y neutro se solapan en ninguno de los casos, cabe destacar que el intervalo de confianza del pitch del tercer segmento de alegría seleccionado no se solapa con los intervalos de los otros dos segmentos como cabría esperar. Este resultado se trasladará al segundo test como veremos en la siguiente tabla.

En este segundo test ninguno de los intervalos contiene el cero, por lo que la diferencia de las distribuciones es estadísticamente significativa entre todos los segmentos de alegría y todos los segmentos neutros seleccionados.
Lo que si observamos, es que los dos límites del intervalo del tercer segmento de alegría son notoriamente inferiores en el caso del pitch si los comparamos con aquellos obtenidos al realizar el test con el primer y segundo segmento de alegría y los dos segmentos neutros. Esto significa que la diferencia de las medias del pitch entre en tercer segmento de alegría y los segmentos neutros es menor que la diferencia de los otros dos segmentos de alegría y los segmentos neutros.

2º test [Z_+, Z.]	Pitch (Hz)		
	1º seg. A	2º seg. A	3º seg. A
1º seg. N	[107,37 , 89,15]	[135,52 , 118,73]	[39,50 , 25,36]
2º seg. N	[90,57 , 89,15]	[118,47 , 106,46]	[21,71 , 13,83]

2º test [Z_+, Z.]	Intensidad (dB)		
	1º seg. A	2º seg. A	3º seg. A
1º seg. N	[8,70 , 5,74]	[11,78 , 8,75]	[9,88 , 6,85]
2º seg. N	[7,20 , 5,65]	[10,28 , 7,07]	[8,38 , 5,18]

Tabla 3. *Resultados del segundo test en los distintos segmentos de alegría y neutro*

Curiosamente este tercer segmento fue clasificado con un 3 (el nivel máximo de alegría) y los dos con un 2, si nos fijamos en la transcripción de cada segmento vemos que el primero y el segundo corresponden a *ah vale* y el tercero a *ole*. Mientras la palabra *ole* si que denota alegría *ah vale* suele usarse más comúnmente para expresar sorpresa, por lo que lo que nosotros habíamos etiquetado como alegría puede que esté más relacionado con la sorpresa que con alegría. Es posible que esa sea la razón de que el valor medio del pitch en estos dos casos sea superior, ya que no son segmentos puramente alegres. En la siguiente figura se muestran tres distribuciones de pitch distintas que encontramos.

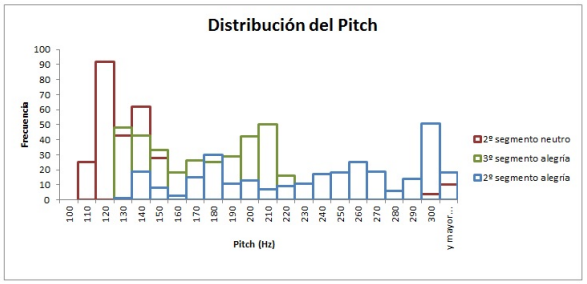

Figura 21. *Distribuciones de pitch del 3 y 2 segmento de alegra y 2 segmento de neutro*

4.2. Primera grabación con enfado

Es una llamada larga de casi 8 minutos de duración y con bastante ruido realizada por una operadora del servicio técnico de Euskaltel a un cliente que se encuentra en China. Esta llamada se realiza para preguntar sobre un problema técnico que lleva teniendo desde que está en China y que no han podido solucionar. El cliente se muestra aun más enfadado cuando la operadora no le puede asegurar que la llamada que le están realizando sea gratuita.

Al contrario que la llamada anterior, ésta se desarrolla en un tono de enfado con pocas expresiones neutras que aparecen sobre todo al principio de la conversación y que por el tono serio del orador se pueden confundir con un enfado moderado.

Los segmentos neutros en este caso son muy diferentes a lo que hemos visto en el caso anterior. El pitch aunque también es más o menos constante no dibuja una línea continuada con pequeñas subidas y bajadas al final de la frase dependiendo de la entonación, sino que son trazos que se cortan abruptamente al final de cada sílaba. La intensidad sin embargo, si que guarda más similitud con las características descritas anteriormente, pero con la diferencia de que en vez de disminuir al final de la palabra disminuye al final de cada sílaba. Ésto junto con el aspecto tan brusco del pitch hace que se la voz se perciba como seca. Nos encontramos ante un discurso más lento que el anterior donde la velocidad en los segmentos neutros varía entre los 11-16 fonemas por segundo. En lo que respecta a las pausas, se observa que aunque si están presentes su duración es muy corta, menor a medio segundo.

Lo más llamativo del pitch en los segmentos de enfado es que aunque siguen cortándose bruscamente, la distribución no es constante, sino que tiene subidas y bajadas. Estos cambios están relacionados con el énfasis que pone el cliente en la palabra que desea subrayar. En lo que respecta a la media es notablemente más alta como se predice para una emoción de activación elevada como es el enfado. La intensidad por su parte sigue el patrón establecido en los segmentos neutros, por lo que tendremos que ver si los clasificadores muestran que existe algún rasgo que permita calificarlo como segmento de enfado o por el contrario en este caso no existe correlación con la emoción y la intensidad de la voz. La velocidad de los segmentos en los que el enfado tiene los niveles más altos es notoriamente superior, rondando los 22 fonemas por segundo llegando hasta los 27. Se observa también un aumento del número de pausas y su duración que en este caso superan el medio segundo por norma general.

En la Figura 22 se muestran las comparativas entre el pitch e intensidad de un fragmento neutro y un fragmento de gran enfado. Ambos tienen una duración similar y cinco sílabas. Las lineas azules corresponden al fragmento de enfado y las verdes al neutro, las líneas discontinuas azules y verdes corresponden a los valores medios obtenidos en el fragmento de enfado y en el neutro respectivamente.

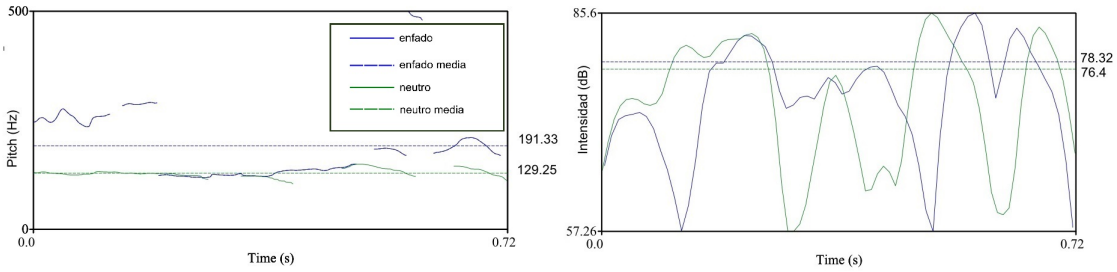

Figura 22. *Distribución del pitch y la intensidad en segmentos neutro* "decisión fue"*, y enfado* "obviamente no"

Al realizar el examen estadístico para el enfado y el estado neutro los resultado que obtenemos son los siguientes:

	Pitch (Hz)	
	Enfado	Neutro
μ	193,36	152,24
σ	82,17	77,05
1º test [Z_+ , Z_-]	[198,90 , 187,82]	[157,94 , 146,55]
2º test [Z_+ , Z_-]	[49,01 , 33,21]	

	Intensidad (dB)	
	Enfado	Neutro
μ	59,83	58,40
σ	16,53	14,51
1º test [Z_+ , Z_-]	[60,70 , 58,96]	[59,08 , 57,72]
2º test [Z_+ , Z_-]	[2,53 , 0,33]	

Tabla 4. *Resultados de los test al comparar el enfado con el estado neutro, los número rojos indican el intervalo el cual está próximo a contener el cero.*

La Tabla 4 muestra que en lo que respecta al pitch tanto el valor medio como la desviación típica aumentan en el enfado, de hecho existe una diferencia de 30 dB en la distribución de las medias de cada uno, el primer test lo respalda puesto que los dos intervalos no se solapan. Por último en el intervalo del segundo test no contiene el cero, por lo que las distribuciones son significativamente diferentes. En la intensidad ocurre algo muy diferente. La figura 22 muestra que la distribución de la intensidad en el estado neutro y enfadado son muy similares. Esto también se refleja al analizar varios segmentos neutros y de enfado. Aunque la media y la desviación típica son ligeramente superiores, los intervalos obtenidos por el primer test se solapan, lo que significa que la distribución de las medias es similar. La diferencia entre las distribuciones tampoco es estadísticamente significativa, puesto que el intervalo obtenido por el segundo test está muy próximo al cero y habrá que hacer el test F para ver si se descarta la hipótesis nula. Este test muestra que para el caso donde se comparan los tres segmentos de enfado y los tres neutros $F_{prueba} = 1,30$ y $F_{tabla} = 3,85$ en este caso. Por lo que $F_{prueba} < F_{tabla}$ consecuentemente se acepta la hipótesis nula, no siendo las dos distribuciones significativamente diferentes.

La tabla 5 muestra que los resultados obtenidos tras el primer test. En ella se ve que los intervalos obtenidos con respecto al pitch no se solapan en general, a excepción del primer segmento de enfado con el tercer segmento neutro que se solapan ligeramente. En la intensidad, por el contrario, lo que domina es el solapamiento de intervalos, por lo que la distribución de las medias en la intensidad tiende a ser similar. La excepción la marca el segundo segmento de enfado, cuyos intervalos tanto de pitch como de intensidad están muy

por encima de los obtenidos en el resto de segmentos, no solo en los pertenecientes a neutro sino que en los pertenecientes a enfado también.

1º test [Z-, Z+]	1º seg. E	2º seg. E	3º seg. E	1º seg. N	2º seg. N	3º seg. N
Pitch (Hz)	[192,04 , 172,87]	[216,6 , 196,30]	[200,26 , 182,39]	[165,11 , 155,96]	[147,89 , 131,70]	[173,67 , 141,13]
Intensidad (dB)	[58,09 , 56,02]	[69,65 , 66,55]	[59,20 , 55,01]	[58,82 , 56,68]	[61,56 , 59,21]	[56,89 , 54,53]

Tabla 5. *Resultados del primer al comparar el enfado con el estado neutro por segmentos, los núeros en azul indican que ocurre solapamiento.*

Los datos de la Tabla 6 indican que existe una tendencia a que las diferencias entre las distribuciones de intensidad no sean estadísticamente significativas, puesto que casi la mitad de ellas contienen el cero en el intervalo obtenido por el segundo test.

2º test [Z-, Z+]	Pitch (Hz)		
	1º seg. E	2º seg. E	3º seg. E
1º seg. N	[32,49 , 11,35]	[5,02 , 34,84]	[40,78 , 20,80]
2º seg. N	[55,14 , 9,44]	[79,57 , 53,74]	[63,52 , 39,53]
3º seg. N	[43,84 , 6,27]	[68,13 , 29,98]	[52,39 , 15,46]

2º test [Z-, Z+]	Intensidad (dB)		
	1º seg. E	2º seg. E	3º seg. E
1º seg. N	[0,78 , -2,18]	[12,22 , 8,47]	[1,69 , -2,99]
2º seg. N	[-1,78 , -2,25]	[9,65 , 5,78]	[-0,90 , -5,67]
3º seg. N	[2,90 , -0,22]	[14,32 , 10,45]	[3,78 , -1,00]

Tabla 6. *Resultados del segundo test al analizar distintos segmentos de enfado y neutros, los números rojos indican que el intervalo contiene el cero*

Haciendo el tercer test para los segmentos que incluyen el cero obtenemos los siguientes resultados:

Test F	1º seg. E		3º seg. E	
	F_{prueba}	F_{tabla}	F_{prueba}	F_{tabla}
1º seg. N	0,94	3,85	1,91	3,85
2º seg. N	1,54	3,85	3,16	3,86

Tabla 7. *Resultados del tercer test*

Por lo que en ningún caso se puede rechazar la hipótesis nula.
Al no poder rechazar la hipótesis nula en estos casos sabemos que la distribución no es significativamente diferente, de hecho si calculamos el hisograma de dos de los segmentos esto es lo que obtenemos:

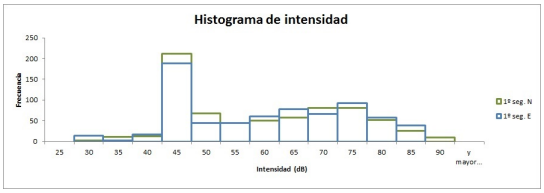

Figura 23. *Comparativa de la distribución de la intensidad del primer segmento neutro y el primero de enfado*

4.3. Segunda grabación con enfado

Esta última grabación de tres minutos de duración es la llamada realizada por un cliente al servicio técnico de Euskaltel, debido a repetidos problemas con la conexión a internet. En este caso el tono general de la llamada también es de enfado. Sin embargo los enfados que aparecen en ambas conversaciones son muy distintos entre si. El primero es un enfado serio y este es un enfado cansado, por lo que es de esperar que las características que analizamos tengan distinto comportamiento.

En lo que al pitch y la intensidad en el estado neutro se refiere, se observa un punto medio entre lo que teníamos en las dos grabaciones anteriores. El pitch aunque no es continuo no se corta en tantas ocasiones como en el caso anterior, tiene una tendencia más continua, lo que significa que el orador no marca tanto las sílabas haciendo una pequeña pausa entre ellas(recordemos que el pitch solo aparece cuando hay voz). Esta continuidad, también se traslada a la intensidad, donde no disminuye tanto al terminar la sílaba pero si al terminar la palabra. La voz por tanto, es más continuada y menos seca que en el caso anterior. La velocidad de locución es ligeramente más rápida que en el caso anterior, puesto que durante la locución etiquetada como neutro se sitúa entre los 16-18 fonemas por segundo. Por lo que respecta al número de pausas se puede decir que coinciden con la puntuación de la frase, lo cual concuerda con lo que cabe esperar de un estado neutro.

En los segmentos donde el enfado es muy notorio se observa que en lo que al pitch se refiere se cumple el patrón del caso anterior y la predicción teórica, es decir las sílabas están más marcadas, los altibajos son evidentes y su valor medio aumenta considerablemente. La distribución de la intensidad por su parte no parece variar mucho de un estado de enfado a uno neutro. Sin embargo sí que vemos como la velocidad aumenta (19-23 fonemas por segundo en los segmentos etiquetados como enfado), tal y como ocurría en el caso anterior. El número de pausas entre los fragmentos etiquetados como enfado es ligeramente superior que en el caso de los fragmentos neutros. Sin embargo no parece que sea una característica relevante a la hora de detectar enfado.

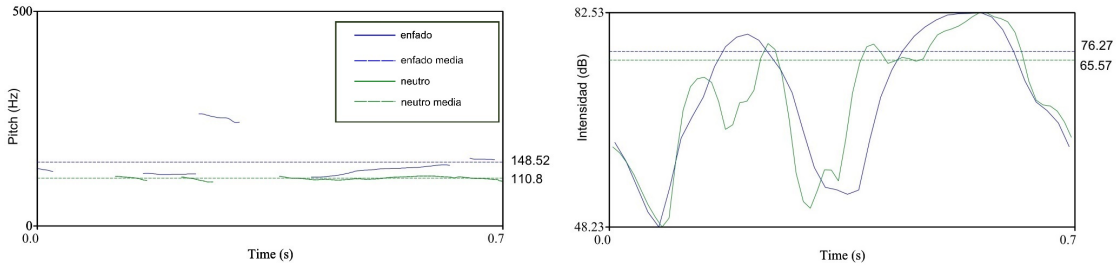

Figura 24. *Distribución del pitch y la intensidad en segmentos neutro y enfado*

Para finalizar examinaremos los resultados obtenidos a través de los test estádisticos.

	Pitch (Hz)	
	Enfado	Neutro
μ	130,21	110,79
σ	44,55	3,33
1º test [Z_+ , Z_-]	[133,52 , 126,90]	[111,24 , 110,35]
2º test [Z_+ , Z_-]	[22,74 , 16,10]	

	Intensidad (dB)	
	Enfado	Neutro
μ	63,17	59,33
σ	12,36	11,15
1º test [Z_+ , Z_-]	[63,70 , 62,64]	[60,16 , 58,50]
2º test [Z_+ , Z_-]	[4,82 , 2,86]	

Tabla 8. *Resultados estádisticos al comparar estados neutros y de enfado*

A diferencia de lo que ocurría en la anterior grabación, en este caso tanto el pitch como la intensidad difieren en sus valores en el estado neutro y enfadado, aunque sí que es verdad que la diferencia del pitch es significativamente mayor. El dato más llamativo es el pequeño valor de la desviación típica en el estado neutro, propiedad que tiene relación con la voz monótona del locutor.

En el análisis por segmentos lo que encontramos es lo siguiente:

1º test [Z_+ , Z_-]	1º seg. E	2º seg. E	3º seg. E	1º seg. N
Pitch (Hz)	[145,23 , 130,72]	[124,70 , 122,67]	[135,49 , 121,74]	[111,24 , 110,35]
Intensidad (dB)	[61,90 , 60,12]	[68,18 , 66,39]	[62,11 , 60,32]	[60,16 , 58,50]

Tabla 9. *Resultados del primer test al analizar distintos segmentos de enfado y neutro, los númers azules muestran que los intervalos se solapan*

El solapamiento entre los intervalos solo ocurre una vez y es prácticamente despreciable. Sí que cabe destacar sin embargo, que en dos de los casos las distribuciones de la probabilidad de las medias en el estado enfadado y neutro están muy juntas, lo que hace pensar que en la intensidad la diferencia de las medias entre estas dos emociones tiende a no ser muy grande. Pero sí que lo es la diferencia entre las medias de pitch.

Al realizar el segundo test esto es lo que obtenemos:

2º test [Z_+ , Z_-]	Pitch (Hz)		
	1º seg. E	2º seg. E	3º seg. E
1º seg. N	[34,41 , 19,95]	[13,99 , 11,80]	[24,68 , 10,97]

2º test [Z_+ , Z_-]	Intensidad (dB)		
	1º seg. E	2º seg. E	3º seg. E
1º seg. N	[2,89 , 0,47]	[9,17 , 6,74]	[3,10 , 0,67]

Tabla 10. *Resultados del segundo test al analizar distintos segmentos de enfado y neutro*

La diferencia entre las distribuciones esta claro que es estadísticamente significativa en el pitch de los segmentos de enfado y neutro. No ocurre así con la distribución de la intensidad que pese a no contener el cero en ninguno de los casos en dos de ellos se aproximan mucho, es decir, las dos distribuciones son similares.

Al realizar el tercer test a los dos segmentos cercanos a cero obtenemos los valores de $F_{prueba} = 1,17$ y 1,14 para el primer y el segundo segmento respectivamente, ambos menores a 3,85 que es $F_{muestra}$ que corresponde a este caso, por lo que la hipótesis nula no puede rechazarse y se deduce que ambas distribuciones son similares.

5. Clasificación y Resultados

En el trabajo de fin de grado de matemáticas consta en utilizar clasificadores capaces de evaluar cada frame y así asignarles la emoción que les corresponde. La clasificación se efectúa en base al pitch, los formantes y la intensidad. Para efectuar la clasificación se utiliza el vector de características del 40 % de los frames de cada grabación como entrenamiento. Como ya hemos comentado en otras ocasiones cada grabación se trata de forma independiente. Para el test se usan el resto de frames y se calcula la probabilidad de éxito al evaluar cada característica por separado. Se utilizaron dos tipos de clasificadores SVM y GMM, el funcionamiento de ambos se explicará brevemente a continuación. [19]

- Modelos de Mezcla Gaussiana (GMM):

  El objetivo de este modelo es parametrizar las características mediante modelos de mezcla gaussianas. Este es un método no discriminatorio. La técnica que se usa se basa en el principio de que cada emoción tiene diferentes sonidos y que la aparición de los sonidos es diferente de una emoción a otra. Los GMM modelan la distribucioón de probabilidad de los parámetros de un fragmento de audio. El modelo total de las mezclas Gaussianas se parametriza mediante el vector de medias, matriz de covarianzas y el peso de mezclas de todas las componentes de densidad.
  Una vez se tenga la configuración y los vectores de entrenamiento de los GMM correspondientes, han de estimarse los parámetros del modelo. Para ello, se usa el método de máxima expectación, uno de los métodos más usados en la práctica.

- Las Máquinas de Vector de Soporte (SVM);

  El objetivo de este clasificador es definir una separación entre las dos clases que queremos clasificar. A diferencia del anterior este es un método discriminatorio y la separacón entre ambas es geométrica. Lo que se busca es establecer una dependencia funcional entre los datos de entrada y los datos de salida. En caso de que la muestra

sea un conjunto separable, el objetivo será maximizar en todo lo posible la distancia para que así sea más fácil diferenciar entre clases. Para casos donde a priori los datos no sean separables, como el nuestro, se puede definir un espacio de más dimensiones donde sí que sean separables. Para lo cual se creará una función b(x) que mapeará el espacio de entrada n-dimensional donde se encuentran nuestros datos, a un espacio de dimensión expandida n'.

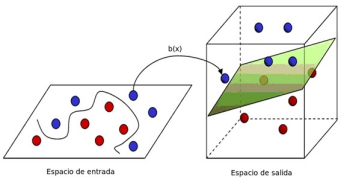

Figura 25. *Mapeo de los vectores bidimensionales a vectores tridimensionales, una explicación más detallada puede encontrarse en [14]*

En el trabajo de mi compañero se usan los SVM basados en supervectores GMM, esta es una técnica de clasificación de patrones que tiene las ventajas de los sistemas parametrizadores, como son los GMMs, y las de los sistemas discriminatorios como son los SVM.

La evaluación de resultados se realizó mediante *Cross Validation* (CV) y *Total Choice* (TC) para garantizar que la partición entre los datos de entrenamiento y prueba son independientes.

Las probabilidades de acierto al evaluar cada característica son las siguientes:

	Alegría				Enfado 1				Enfado 2			
	SVM		SVM_GMM		SVM		SVM_GMM		SVM		SVM_GMM	
	CV	TC	CV	TC	CV	TC	CV	TC	CV	TC	CV	TC
Formantes	0,78	0,81	0,78	0,83	0,76	0,83	0,74	0,75	0,88	0,78	0,82	0,78
Pitch	0,89	0,93	0,78	0,83	0,74	0,8	0,74	0,75	0,82	0,78	0,82	0,78
Intensidad	0,78	0,88	0,78	0,83	0,7	0,72	0,74	0,75	0,82	0,76	0,82	0,78

Tabla 11. *Resultados obtenidos por los clasificadores, extraído de [19]. En verde están resaltados los mejores resultados.*

6. Conclusiones y Trabajo Futuro

La segunda parte de este trabajo se ha centrado en el análisis de tres llamadas reales y en general se puede decir que los resultados coinciden la revisión bibliográfica realizada, aunque siempre hay que tener en cuenta que estos resultados se han obtenido de un número muy pequeño de muestras por lo que no es posible tomarlos como patrón de comportamiento.

La detección de emociones es una tarea complicada, no sólo en el ámbito de la detección automática, sino que la detección de emociones por parte de los humanos también lo es. La percepción de emociones en el habla es algo subjetivo y aunque por lo general para las emociones bien marcadas se espera que la concordancia entre los oyentes sea muy alta no

lo es tanto en casos donde la expresión de la emoción es más sutil o existe una mezcla de emociones. Por otro lado, también hay que tener en cuenta la personalidad del individuo. Es conocido que hay personas que expresan sus emociones de una forma más efusiva que otras, por lo que aunque dos personas estén expresando la misma emoción, dependiendo de la efusividad con la que la expresen puede ser más o menos difícil para un oyente determinar de que emoción se trata.
Por supuesto todos estos problemas se trasladan a la detección autómatica de emociones haciendo la tarea mucho más complicada, puesto que un ordenador carece de ese factor humano que a veces es necesario a la hora identificar una emoción. Sin embargo, parece que las emociones bien marcadas especialmente aquellas de elevada activación, como lo son la alegría y la ira, son más sencillas detectar si se fija la atención en las características más claras y se usan los clasificadores adecuados.
En este proyecto, como se ha mencionado varias veces, se ha trabajado con llamadas reales. Esto supone una ventaja con respecto a usar actores que *interpreten* la emoción, puesto que te da la certeza de que la emoción se expresa de una manera real y no de la manera que un humano espera que la expresión de esa emoción sea. Pero también tiene la desventaja de no tener tantas muestras como se precisen de esa emoción o del estado neutro. Este problema es visible en el caso de las grabaciones de enfado especialmente, puesto que los fragmentos neutros son muy escasos y quedan muchas veces ocultos entre el tono de enfado presente en la llamada. Esto supone que no hay muchas comparaciones que se puedan hacer entre las propiedades de las características de la señal de voz en el estado neutro y en el enfadado. Por lo tanto no se puede saber a ciencia cierta si las características que aparecen cuando se aprecia enfadado son efecto de la expresión de enfado del cliente o por el contrario, guardan relación con su forma de hablar.
Este fenómeno también ocurre en la grabación con alegría pero de forma contraria. En este caso son escasos los fragmentos que muestran alegría y además todos ellos tienen características similares en cuanto a duración y número de palabras. Por lo que lo que hemos observado en estos fragmentos puede que no sea válido para fragmentos de una mayor duración.
En lo relativo a la variación de las características en la emoción de alegría hemos visto como en general la intensidad y el pitch aumentaban su valor medio considerablemente, por lo que parece que si que coincide con lo predicho por otros estudios. Los resultados de a la velocidad no son concluyentes, puesto que los segmentos de alegría eran escasos y de corta duración. La velocidad en los segmentos alegres es menor que en los segmentos neutros, pero no hay muestras suficientes para afirmarlo.

Por lo que respecta al enfado, hemos visto que tiene una activación elevada, por lo que las características en muchos aspectos tendrían que parecerse a lo visto en la expresión de alegría. Esto no ocurre del todo así. En las dos llamadas de enfado es mucho más notorio el aumento de la media del pitch que el de la intensidad. Un detalle muy llamativo es el aspecto entrecortado del pitch durante el enfado. Al parecer el orador mete pequeñas pausas entre las sílabas cuando demuestra enfado. Este es un aspecto puede apreciarse únicamente haciendo este tipo de análisis de la señal siendo inapreciable ante los clasificadores que se usan en el proyecto de fin de carrera de mi compañero.

Al introducir las dos llamadas telefónicas que mostraban enfado dijimos que ambos enfados eran diferentes por lo que se esperaba que la variación del pitch, intensidad o velocidad lo mostrasen de alguna manera. Existen dos diferencias notables entre las dos grabaciones, la primera es que la variación de la media del pitch con respecto al estado neutro es casi el doble en la primera grabación comparándola con la segunda. La segunda es que la di-

ferencia de la desviación típica entre el estado enfadado y el neutro es, por el contrario, muchísimo mayor en el segundo caso. Si comparamos los datos de la desviación típica para ambas grabaciones obtenemos 82.17 dB y 77.05 dB para la primera grabación en el caso de enfado y neutro respectivamente y 44.55 dB y 3.38 dB para la segunda. Por lo que aunque si que es cierto que en el segundo caso la diferencia entre ambas emociones es muy superior, también es verdad que el valor para el enfado en la primera grabación es casi el doble que en la segunda. Esto es consecuencia del tono monótono del segundo cliente, hecho que puede tener relación con el carácter cansado de su enfado.

Los resultados positivos de mi compañero parecen indicar que las características que hemos analizado son relevantes en la detección de emociones en el habla, por lo que a trabajos futuros se refiere sería conveniente seguir trabajando en esta linea. Para poder trabajar con un volumen de muestras más elevado, sería conveniente automatizar el proceso de selección de los segmentos útiles de la señal y extracción de características. Así mismo, sería conveniente utilizar en vez de las características absoluta las diferencias comparándolas con el estado neutro para así poder entrenar un algoritmo general que sirva para todos los individuos. Por último, sería positivo añadir el análisis espectral al proceso de clasificación, puesto que en muchos estudios [14] se cita como herramienta eficaz en este campo.

Gracias a este trabajo he podido no solo adquirir competencias en el ámbito de la detección automática de emociones, sino que ha sido también una oportunidad para utilizar conceptos aprendidos durante la carrera en un problema real. Así mismo me ha servido para aprender a usar distintos programas como Praat, Latex o Excel. También para mejorar mis conocimientos de estadístca. No menos importante han sido las oportunidades de poder ver el funcionamiento de una empresa y colaborar con un compañero de otra titulación.

Los resultados obtenidos tanto en este proyecto como en el realizado por mi compañero serán enviados de manera conjunta al congreso nacional *Iberspeech* que se realizará el próximo noviembre en Las Palmas de Gran Canaria.

Bibliografía

[1] Jacob Benesty, M. Mohan Sondhi y Yiteng Huang *Speech processing*, Springer Handbook.

[2] Antonio Peinado y José Segura *Speech recognition over digital chanels*, Wiley.

[3] J.L: Flanangan *Speech Analysis Synthesis and Perception*, Springer -Verlag

[4] J.D. Markel y A. H. Gray Jr *Linear Prediction of Speech* Springer-Verlag

[5] K. Sreenivasa Rao y Shashidhar G. Koolagudi *Robust Emotion recognition using Spechtral and prosodic features*, Springer.

[6] Monique Adriana Johanna Biemans *Gender variation in voice quality*, LOT 2000

[7] Ricardo Gutierrez-Osuna *Introduction to Speech processing* TAMU

[8] Christer Gobl y Ailbhe N. Chasaide *The role of voice quality in communicating emotion, mood and attitude* Speech comunication

[9] Ioulia Grichkovtsovaa,Michel Morel y Anne Lacheret *The role of voice quality and prosodic contour in aective speech perception* Speech comunication

[10] Alan Ó Cinnéide *Linear Prediction, the Technique, its Solution and Application to Speech* Dublin Institute of Technology, Institiúid Teicneolaíochta Átha Cliath

[11] K. Ishizaka y J.L. Flanagan*Synthesis of voiced Sounds From a Two-Mass Model of the Vocal Cords*, The Bell System Technical Journal 1972

[12] Gonville y Caius College *Emotion Detection from Speech*, Computer Science Tripos Part II.

[13] Laurence Vidrascu y Laurence devillers *Detection of real-life emotions in call centers*, Interspeech.

[14] Carlos Ortego Resa *Detección de emociones en voz espontánea*, Universidad Autonoma de Madrid.

[15] Moataz el Ayadi, Mohamed S. Kamel y Fakhri Karray *Survey on speech emotion recognition Features, classification schemes and databases*, Pattern Recognition.

[16] Humperto Pérez Espinosa y Carlos Alberto Reyes García *Reconocimiento de Emociones a Partir de Voz Basado en un Modelo Emocional Continuo*, Coordinación de Ciencias Computacionales INAOE.

[17] Björn Schuller, Anton Batliner, Stefan Steidl y Dino Seppi *Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challange*, Science Direct.

[18] Shashidhar G. Koolagudi y K. Sreenivasa Rao *Emotion recognition from speech: a review*, Springer.

[19] Manex Serras *Deteccin y clasificacin de parmetros de identificacin de habla emocional*, Universidad del País Vasco/ Euskal Herriko Unibertsitatea

[20] *http://opensmile.sourceforge.net*

[21] *http://haroon.99k.org/page16.html*

[22] *http://www.fon.hum.uva.nl/praat/*

[23] *http://www.montgomerycollege.edu/Departments/StudentJournal/Automatic.pdf*

[24] *http://www.stanford.edu/class/linguist156/May26_therest.pdf*

A. Apéndice: Reuniones y Actividades

Para la realización de este proyecto se han producido las siguientes reuniones:

- 11.09.2013 Reunión con la empresa Xupera
- 28.10.2013 y 11.11.2013 Seminario de defensa efectiva del trabajo de fin de grado
- 11.11.2013 Reunión para organizar un plan de trabajo
- 18.12.2013 Presentación de la primera parte del trabajo ante el grupo de investigación
- 14.02.2014 Reunión para programar la segunda parte del trabajo
- 13.03.2014 Reunión para el análisis del etiquetado
- 21.05.2014 Reunión para compartir resultados y decisiones
- 17.06.2014 Reunión para la evaluación de resultados