PROYECTO FIN DE CARRERA

Ingeniero Técnico de Telecomunicación

Especialidad Sistemas de Telecomunicación

Similitud y confusión fonéticas: reconocimiento humano vs. reconocimiento automático mediante técnicas entrópicas y de análisis de conceptos formales

Autor: Sira González Martín

Tutor: Carmen Peláez Moreno

Co-director: Francisco José Valverde Albacete

Leganés, Julio de 2014

Universidad Carlos III de Madrid

Ingeniería Técnica de Telecomunicación: Sistemas de Telecomunicación

Título: Similitud y confusión fonéticas: reconocimiento humano vs. reconocimiento automático mediante técnicas entrópicas y de análisis de conceptos formales

Autora: Sira González Martín.

Tutora: Carmen Peláez Moreno.

Co-director: Francisco José Valverde Albacete.

EL TRIBUNAL

Presidenta: Ascensión Gallardo Antolín

Secretario Rubén Solera Ureña

Vocal: Luis Antonio Puente Rodríguez

Realizado el acto de defensa y lectura del Proyecto Fin de Carrera el día _23_ de _____julio_____ de __2014__ en Leganés, en la Escuela Politécnica Superior de la Universidad Carlos III de Madrid, acuerda otorgarle la CALIFICACIÓN de:

AGRADECIMIENTOS

Agradezco a todas las personas que durante estos años han formado parte de estos momentos, a mis padres, hermana, compañeros durante estos años de esfuerzo y a mis amigos por estar ahí en todo momento y especialmente a mis tutores, por su dedicación, paciencia y apoyo. A todos vosotros gracias.

RESUMEN

En el presente proyecto se analiza y compara el comportamiento en reconocimiento del habla de una persona frente a un reconocedor automático del habla híbrido.

El objetivo de nuestro estudio es el de analizar las características acústico-articulatorias que marcan los factores de reconocimiento humano, de esta manera seremos capaces de mejorar los reconocedores de habla automática.

Dicho estudio se ha llevado a cabo haciendo uso de herramientas novedosas para el análisis de los errores de un reconocedor. Estas herramientas son los triángulos entrópicos, con los que medimos la calidad de un reconocedor con otras figuras de mérito que el valor de la precisión, y el análisis formal de conceptos que nos permitirá representar, para su posterior análisis, las confusiones de las bases de datos usadas mediante retículos de confusión.

Gracias a estas herramientas tendremos una visión más clara de las diferencias en la percepción de las características acústico-articulatorias entre el reconocedor de habla humano y el automático. De este modo podremos sugerir nuevas líneas de investigación para mejorar dichas características en los reconocedores.

ABSTRACT

In this project we analyze and compare the performance in speech recognition of a person facing a hybrid automatic speech recognizer.

The aim of our study is to analyze the acoustic-articulatory factors that marks the human recognition, so we will be able to improve automatic speech recognizers' features.

This study was carried out by using new tools for analyzing errors in automatic speech recognizers.

These tools are entropic triangles, with which we measure the quality of a recognizer with other figures of merit that the value of accuracy, and formal analysis of concepts that allow us to represent, for further analysis, confusions bases data used by lattices of confusion.

With these tools we will have a clearer view of the differences in the perception of acoustic-articulatory features between human and automatic speech recognizer. Thus we can suggest new lines of research to improve these characteristics recognizers.

.

ÍNDICE

Índice de figuras

1 INTRODUCCIÓN

1.1 MOTIVACIÓN

Si en algo se ha diferenciado a lo largo de los siglos, y se sigue diferenciando, el ser humano de otras especies es de la capacidad de comunicarse mediante la utilización el lenguaje, ya sea oral o escrito. Gracias al lenguaje el ser humano ha avanzado a pasos agigantados siendo capaz de evolucionar y crear tecnología impensable en otros tiempos.

Gracias a los avances tecnológicos asociados a la información, nuestra sociedad está cada día más conectada electrónicamente. Por este motivo labores que tradicionalmente eran realizadas por seres humanos son ahora realizadas por sistemas automatizados.

Entre esos sistemas automatizados se encuentran aquellos cuyo objetivo es el de que dicho sistema sea capaz de reconocer el habla humana: es lo que llamamos reconocedor de habla automática (RAH o en inglés ASR, "Automatic Speech Recognizer").

El reconocimiento de habla automático es un problema tecnológico cuya solución está lejos de alcanzar las prestaciones de los humanos en esta tarea que denominaremos Reconocimiento Humano del Habla (RHH o en inglés, HSR, "Human Speech Recognition").

En los resultados de un reconocedor de habla automático influyen muchos y muy variados factores lo que dificulta el proceso de entrenamiento y evaluación. Entre estos factores se encuentra cuál es el procedimiento para representar las características acústico-articulatorias utilizado para presentar las señales de voz al reconocedor, es decir, el método de extracción de características. Si este procedimiento es capaz de capturar características que permiten discriminar entre las diferentes unidades acústicas el número de errores de reconocimiento será pequeño. Si, por el contrario, este procedimiento no es discriminativo, el reconocedor no será capaz de hacer la decodificación acústica.

En este proyecto adoptamos la postura de que para poder construir buenos reconocedores primeramente debemos ser capaces de saber cuáles son las características acústicas o articulatorias que marcan los factores de reconocimiento humano, es decir, para mejorar los reconocedores automáticos debemos saber cuáles son las circunstancias bajo las cuales se produce la confusión (acústicamente hablando) en los humanos para así poder diseñar reconocedores más eficaces.

La motivación de la realización de este proyecto es investigar más en este ámbito y poder dar directrices a seguir para mejorar el funcionamiento de los reconocedores mediante representaciones novedosas de la calidad de un reconocedor y el estudio de las confusiones articulatorias del mismo.

1.2 OBJETIVOS

Nuestro objetivo a lo largo las siguientes páginas es estudiar y comparar el comportamiento de lo que llamamos "reconocedor de habla humano" con el "reconocedor máquina" centrándonos en las características acústico-articulatorias del habla.

Para ello tomaremos como referencia los experimentos sobre reconocimiento humano del habla realizados por Miller & Nicely [1]. Por otro lado replicaremos estos experimentos sobre una base de datos limpia y luego mezclada con ruido para ver cuál es el comportamiento de un reconocedor automático en la misma situación en la que Miller & Nicely colocaron a los sujetos humanos de su experimento.

En el presente documento describimos las herramientas innovadoras que utilizamos para la interpretación de la matriz de confusión y las medidas de calidad del reconocedor que nos proporcionarán la información necesaria para poder obtener conclusiones y líneas futuras de investigación.

Dichas herramientas serán los conocidos como "triángulos entrópicos" (ET, del inglés: "Entropy Triangle"), que son capaces de medir la calidad de un reconocedor a partir de una matriz de

confusión sin reducirla exclusivamente a la información de la diagonal como hace la medida clásica de precisión (*accuracy*, en inglés). Esto nos permitirá obtener información más interesante sobre los casos en los que el reconocedor se comporta correctamente y los casos en los que no lo hace, como un primer paso a mejorar los mismos.

La siguiente herramienta es el análisis formal de conceptos generalizado que, mediante la representación de los retículos de confusión, nos ayudará a interpretar las confusiones en los reconocedores mediante una técnica de análisis exploratorio de datos, de manera que podamos representar las relaciones y características lingüísticas que provocan la confusión.

1.3 ORGANIZACIÓN DEL DOCUMENTO

El presente documento está organizado en un total de 5 capítulos, que a continuación vamos a pasar a listary ofrecer un pequeño resumen de los mismos:

- Capítulo 1: nos ofrece una introducción al reconocimiento del habla y nos presenta el objetivo del documento.
- Capítulo 2.: expone las bases teóricas del reconocimiento del habla (tanto humano como automático).
- Capítulo 3: nos presenta las bases teóricas de las distintas herramientas para el análisis de los errores de los reconocedores.
- Capítulo 4: describe los experimentos realizados, las bases de datos usadas además del análisis y representación de los resultados obtenidos.
- Capítulo 5: recoge las conclusiones obtenidas de nuestro estudio y las líneas futuras de actuación.

2 RECONOCIMIENTO DEL HABLA HUMANO Y AUTOMÁTICO.

2.1 RECONOCIMIENTO HUMANO DEL HABLA: MILLER & NICELY, 55

Lo que conocemos como reconocimiento humano del habla (RHH) es la acción de la comunicación hablada entre dos personas, una emite y la otra recibe e interpreta dicha habla.

Para nuestro estudio, tomaremos como ejemplo de RHH los experimentos llevados a cabo en 1955 por los George A. Miller y Patricia Nicely [1].

Éstos, habiendo estudiado el comportamiento del habla durante años intentaron demostrar que muchos de los errores de reconocimiento del habla seguían un patrón.

Partiendo de esta base, pensaban que se podrían aprender patrones sobre la percepción del habla y de esta manera ser capaces de mejorar las máquinas de reconocimiento del habla sabiendo de antemano cuales son los errores más comunes, y así buscar maneras de corregirlos.

Los experimentos que llevaron a cabo (véase apartado 4.1.1) se basaban en la locución por parte de humanos de un pequeño conjunto de sílabas sin sentido y el reconocimiento de las mismas por parte de los mismos humanos que participaban por turnos como locutores.

En los experimentos se jugó con la SNR (relación señal a ruido) y las frecuencias de corte de las señales de voz, para ver cómo se comportaban los reconocedores en situaciones adversas y así poder sacar conclusiones de sus hallazgos.

Cabe destacar que en sus experimentos el hecho de que las sílabas emitidas carezcan de significado evita que los sujetos del test hagan uso de sus conocimientos lingüísticos para identificar los sonidos. Veremos más adelante que el correlato de este conocimiento en el caso automático es el denominado "modelo de lenguaje" del cual prescindiremos en este proyecto por los mismos motivos que lo hicieron Miller & Nicely[1] en su día.

2.1.1 EL ANÁLISIS CON RUIDO

Con el fin de estudiar y analizar el comportamiento de un reconocedor frente al ruido Miller & Nicely modificaron la SNR en su base de datos y replicaron su experimento inicial con las siguientes SNR:

[-18db, -12db,-6db, 0db, 6db, 12db]

Cabe destacar que las conclusiones más interesantes de los experimentos realizados se obtuvieron del análisis con supresión de bandas frecuenciales.

2.1.2 EL ANÁLISIS CON SUPRESIÓN DE BANDAS FRECUENCIALES

Miller & Nicely modificaron el espectro de frecuencia de las señales emitidas para estudiar el comportamiento del reconocimiento humano del habla, mediante el filtrado de diferentes bandas de frecuencia..

Podemos dividir las señales en las filtradas paso alto y las paso bajo, siendo las frecuencias de corte analizadas las siguientes:

Paso bajo:

Frecuencia corte inferior	Frecuencias corte superior
200	[300,400,600,1200,2500,5000,6500]

Paso alto:

Frecuencias corte inferior	Frecuencia corte superior
[1000,2000,2500,3000,4500]	5000

[1] En lo sucesivo nos referiremos a estos experimentos como M&N55

De esta manera obtuvieron 12 matrices de confusión. En estas matrices de confusión agregadas de todos los locutores podemos ver el comportamiento idealizado de un locutor humano que pone en juego exclusivamente su desempeño fonético. En este proyecto replicaremos estos experimentos obteniendo resultados equivalentes para reconocedores automáticos.

Las matrices por sí solas no nos dan información de interés de cómo se confunden unas sílabas con otras así que Miller & Nicely decidieron dividirlas en pequeños subconjuntos, para poder estudiarlos. Para ello se basaron en un modelo de 5 canales de transmisión que ellos mismos idearon y que pasaremos a explicar en el apartado 2.1.3

2.1.3 CONCLUSIONES DE LOS EXPERIMENTOS DE M&N55

Durante años, la clasificación de los fonemas se ha hecho teniendo en cuenta distintas características del proceso de articulación para generar sonidos.

Estas características de la producción del habla se reflejan en ciertas características acústicas que son presumiblemente discriminadas por el receptor.

Cuando Miller & Nicely realizaron sus experimentos se dieron cuenta que con las matrices de confusión completas no veían con claridad lo que pretendían analizar así que decidieron subdividir los resultados agrupándolos de tal manera que fuesen interesantes para sus propósitos. Este proceso lo llevaron a cabo mediante la observación de las matrices de confusión generadas realizando permutaciones en filas y columnas hasta conseguir que las matrices se aproximaran a matrices diagonales por bloques, aunque carecían de un procedimiento algorítmico para esta reorganización.

De este modo llegaron a la conclusión de que lo más natural es guiarse por las características articulatorias determinando que, para las 16 consonantes estudiadas, existían las siguientes cinco características pertinentes para la clasificación de los resultados obtenidos.

Sonoridad (Ing. “Voicing”).

Fonéticamente hablando podemos distinguir los fonemas sonoros (aquellos que al pronunciarse hacen que nuestras cuerdas vocales vibren) y los sordos (aquellos que al pronunciarse no hacen vibrar nuestras cuerdas vocales). Se dividirán las consonantes estudiadas teniendo en cuenta esa característica.

Nasalidad (Ing. “Nasality”).

Para pronunciar los sonidos /m/ y /n/ utilizamos la cavidad nasal, es por ello que introducimos una diferencia acústica respecto a la forma de pronunciar otros fonemas. Además tenemos que tener en cuenta que estas son las dos únicas consonantes que no tienen la componente aperiódica del ruido.

Fricción o fricación (Ing. “friction”)

Con esta característica lo que generamos es un tipo de sonido consonántico obstruyente que se inicia con una oclusión (obstrucción del flujo de aire) y una fricación (liberación del flujo de aire) de forma rápida y sucesivamente entre los órganos articulatorios **¡Error! No se encuentra el origen de la referencia.**,[3] generando lo que denominamos "consonantes africadas"**.** A estas consonantes también se les considera como una combinación de una consonante oclusiva y una fricativa.

Duración (Ing. "Duration")

Es el nombre que Miller & Nicely crearon arbitrariamente para designar la diferencia entre /s/, /ʃ /, /z/, /ʒ/ y las otras doce consonantes. Estas cuatro consonantes son más largas, intensas, con más ruido a altas frecuencias, pero en su opinión es su duración extra lo que resulta más efectivo para definirlo como característica.

Punto de articulación (Ing, "Place of articulation")

Esta característica está ligada al uso de la cavidad vocal a la hora de pronunciar los fonemas; es el punto en el que se produce el contacto entre la lengua y la cavidad vocal a la hora de pronunciar un sonido.

Usualmente se consideran tres posiciones: frontal, media y posterior.

En los experimentos se clasificaron las consonaste de la siguiente manera:

- Frontal: /p/, /b/, /f/, /v/, /m/
- medio: /t/,/d/, /θ/, /s/, /ð/, /z/, /n/
- posterior: /k/, /g/, /ʃ /, /ʒ/

Aunque estas tres posiciones son fáciles de reconocer en la producción de esos sonidos, las consecuencias acústicas de cambiar el punto de articulación pueden ser mucho más complejas.

Miller & Nicely intentaron analizar los resultados obtenidos considerando únicamente estas características y haciendo medidas de covarianza llegaron a las siguientes conclusiones:

a) Podemos aplicar la medida de la covarianza a las distintas características lingüísticas de forma separada. De esta manera procederemos a dividir la matriz de confusión resultante en cinco matrices más pequeñas cada una de ellas asociadas a una de las cinco características.

b) Si lo analizamos así nos damos cuenta que es equivalente a considerar que en realidad estamos probando cinco **canales** de comunicación diferentes de forma simultánea[4].

Por supuesto, estos cinco canales no serán independientes, aunque el "cross-talk" será tan pequeño que las características se percibirán casi independientemente unas de las otras.

c) El hecho de que las medidas realizadas para los diferentes canales puedan sumarse de forma simple para obtener un valor aproximado del total de las sílabas transmitidas es bastante indicativo de su independencia.

d) Esta independencia perceptual de las distintas características implica que todo lo que debemos saber sobre un sistema es cuán bien transmite las "indicios acústicos" necesarios para cada característica; las medidas para una característica individual pueden hacerse más fácilmente y rápidamente que la medida para el canal entero, y el factor de corrección para las entradas redundantes depende enteramente del vocabulario de entrada y no de los test de cada experimento.

2.2 RECONOCIMIENTO AUTOMÁTICO DEL HABLA

2.2.1 FUNDAMENTOS

El Reconocimiento Automático del Habla (RAH o en inglés ASR, "Automatic Speech Recognition") tiene como objetivo facilitar la comunicación hablada entre seres humanos y máquinas. Un sistema de reconocimiento de voz es una herramienta capaz de procesar la señal de voz emitida por el ser humano y reconocer la información contenida en ésta, convirtiéndola en texto, es decir, **transcribiéndola**.

El problema que se plantea en un sistema de RAH es el de hacer cooperar un conjunto de informaciones que provienen de diversas fuentes de conocimiento lingüístico (acústica, fonética, fonológica, léxica, sintáctica, semántica y pragmática), en presencia de ambigüedades, incertidumbres y errores inevitables, para llegar a obtener una interpretación aceptable del mensaje acústico recibido [5].

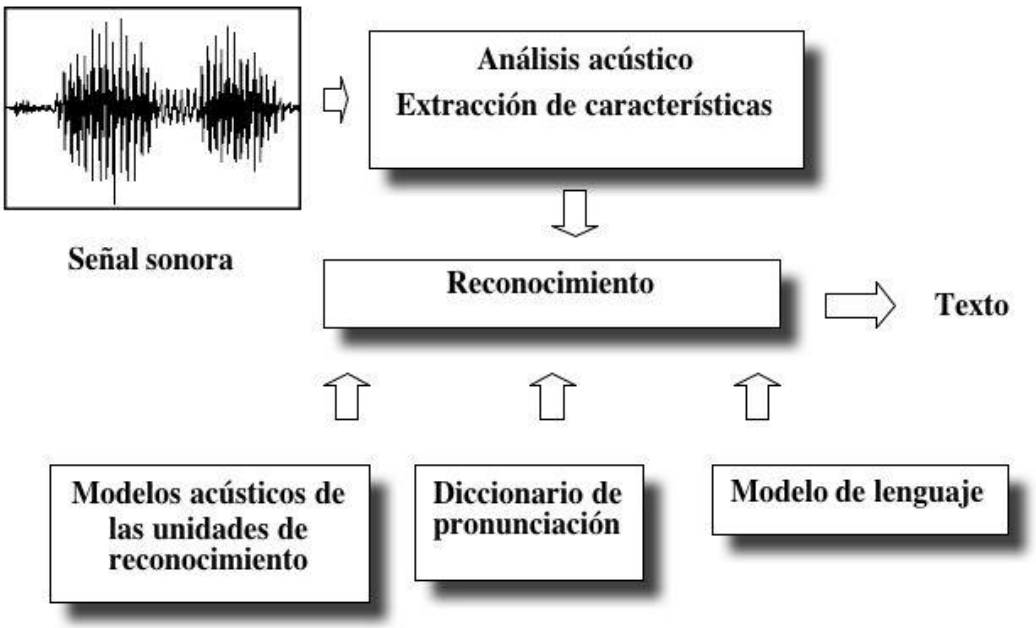

Figura 1: diagrama de bloques de un sistema de RAH, de [6]

Tal y como se puede ver en la Figura 1, en un primer momento, un Sistema de Reconocimiento Automático de Habla captura la señal acústica emitida y la transforma a un formato digital. Después la parametriza, es decir, extrae de la señal de voz sus características más representativas, reduciendo así el tamaño y la variabilidad propias de las ondas acústicas, y convirtiéndola en un vector de parámetros. Finalmente realiza la transcripción utilizando para ello los denominados "Modelo Acústico" y "Modelo de Lenguaje". A continuación vamos a presentar los bloques anteriores haciendo especial énfasis en los que atañen a nuestro sistema.

2.2.2 EXTRACCIÓN DE CARACTERÍSTICAS

Las técnicas de pre procesado o extracción de características más usuales en los sistemas ASR son los MFCC (Ing. "Mel Frequency Cepstral Coefficients", o Coeficientes Cepstrales en Escala Mel) y los Coeficientes PLP (Ing. "Perceptual Linear Prediction", o de Predicción Lineal Perceptual). A continuación los definiremos de forma somera puesto que no son el objetivo de nuestro proyecto, aunque obtendremos ambos coeficientes en la realización de nuestros experimentos[7].

2.2.2.1 Coeficientes MFCC:

La combinación del análisis cepstral de la señal de voz y la noción de una transformación de la escala lineal de frecuencias en función de la influencia que poseen las bandas críticas en la sensibilidad del sistema auditivo humano, da lugar a la técnica de análisis denominada "Mel Cepstrum". Los coeficientes obtenidos mediante este tipo de análisis, son los MFCC.

Estos coeficientes representan el habla basándose en algunos aspectos de la percepción auditiva humana, con la ayuda de diferentes técnicas, como son la de la Transformada de Fourier para obtener las componentes frecuenciales de la señal, así como de un Banco de Filtros Mel que modelan la respuesta de la cóclea espaciando las bandas de frecuencias de manera logarítmica. Posteriormente se aplica una DCT (Ing. "Discrete Cosine Transform", o Transformada Discreta del Coseno) para compactar la información y poder realizar la separación entre la envolvente espectral (procedente del tracto vocal) y la excitación mediante un "liftering".

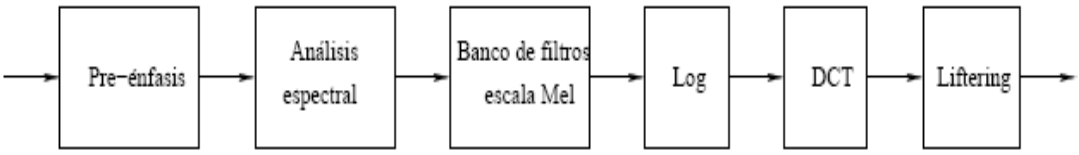

Figura 2: Diagrama de bloques general de la obtención de los parámetros MFCC.

2.2.2.2 Coeficientes PLP:

Otra de la parametrizaciones más utilizadas en reconocimiento de voz, es la Predicción Lineal Perceptual, que comparte muchos aspectos con MFCC. En este caso el filtrado por bandas Mel se sustituye por un análisis de bandas críticas, que aunque modela de manera distinta las bandas de frecuencia, sigue teniendo por objetivo el de emular las diferentes sensibilidades del oído a altas y bajas frecuencias, en este caso añadiendo el efecto auditivo conocido de que las diferentes bandas frecuenciales poseen distintos umbrales de percepción y se perciben las amplitudes de forma distinta.

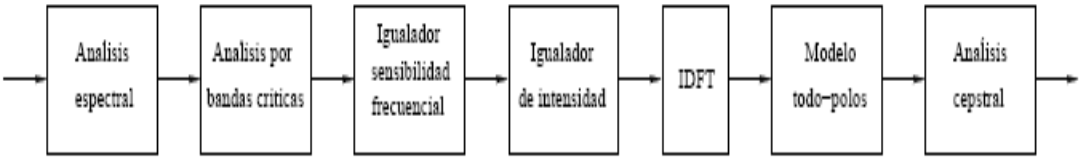

Figura 3: Diagrama de bloques general de la obtención de los parámetros PLP.

2.2.3 MODELADO ACÚSTICO

Una vez obtenidos las características de la voz, los modelos acústicos en el reconocimiento automático del habla tienen la tarea de entrenar un modelo que simularía la señal de voz original (modelos generativos) para así poder decidir qué es lo que ha emitido el emisor y transcribirlo a texto (con la ayuda de un modelo de lenguaje).

Esto le hace desempeñar un papel fundamental en la mejora de la precisión y podría decirse que es la parte central de cualquier sistema de reconocimiento de habla.

Existen distintos tipos de modelos acústicos para el reconocimiento automático del habla, en nuestro caso vamos a citar y explicar brevemente algunos de los más importantes.

Tendremos en cuenta que el reconocedor de habla utilizado en este proyecto estará basado en uno de ellos: reconocimiento híbrido de redes neuronales y modelos ocultos de Markov.

2.2.3.1 Modelos Ocultos de Markov (HMM: Hidden Markov Model)

Los modelos ocultos de Markov[8] se han convertido en uno de los métodos estadísticos más potentes para modelar las señales del habla. Sus principios han sido utilizados con éxito en reconocimiento de voz automático, análisis de tonos y formantes, mejora de expresión, síntesis de voz, comprensión de lenguaje hablado [9] etc.

Un HMM es una máquina de estados finitos, en la que las observaciones son una función probabilística del estado. Es decir, el modelo es un proceso doblemente estocástico formado por un proceso estocástico oculto no observable directamente, que corresponde a las transiciones entre estados y un proceso estocástico observable cuya salida es la secuencia de vectores espectrales.

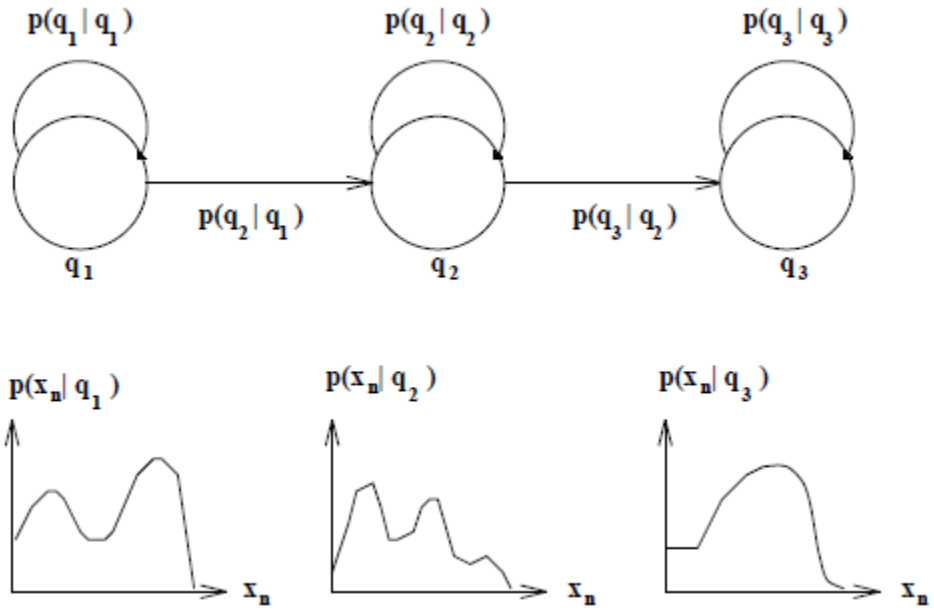

Figura 4: Ejemplo de transición de estados en un modelo oculto de Markov

Cuando trabajamos con HMM nos encontramos con tres problemas básicos que se deben resolver:

1. *Problema de Evaluación:* Dada una secuencia de observaciones y un modelo, se busca cómo calcular la probabilidad de que la secuencia observada haya sido producida por dicho modelo.

2. *Problema de Decodificación*: Dada una secuencia de observaciones y un modelo, se busca cómo elegir una secuencia de estados que sea óptima en algún sentido para producir la secuencia de observaciones.

3. *Problema de aprendizaje*: Dada una secuencia de observaciones de entrenamiento, se busca cómo obtener los parámetros del modelo de forma óptima.

Como hemos visto anteriormente los HMM tienen multitud de aplicaciones. Centrándonos en la que nos compete que es el reconocimiento de voz, podemos aplicar qué modelo HMM representaría una unidad acústica (fonema, difonema, trifonema o cualquiera otro) donde el número de nodos dependerá de la elección de dicha unidad acústica; cada uno de los nodos tiene asociada una probabilidad de emisión; cada sucesión de fonemas con su probabilidad de transición asociada se agrupará en unidades de orden superior (palabras, frases,…), etc.

2.2.3.2 Redes Neuronales artificiales

Las redes neuronales artificiales (RNA, Ing. “Arficial Neuronal Networks”, ANN) son sistemas de procesamiento de información cuya estructura y funcionamiento están inspirados en las redes neuronales biológicas. Desde que el psicólogo Frank Rosenblatt en 1957 introdujo el modelo del perceptrón de una sola capa, las RNA se convirtieron en una herramienta poderosa para solucionar diversos tipos de problemas relacionados con la clasificación, estimación funcional y optimización del reconocimiento de patrones.

En todo modelo de RNA se tienen cuatro elementos básicos:

1. Un conjunto de conexiones, pesos ó *sinapsis*, que determinan el comportamiento de la neurona, las cuales pueden ser *excitadoras*, si presentan un signo positivo (conexiones positivas) y o *inhibidoras,* si presentan un signo negativo (conexiones negativas).
2. Una *función de combinación,* que se encarga de sumar todas las entradas multiplicadas por sus pesos correspondientes.
3. Una *función de activación* que puede ser lineal ó no lineal, empleada para modelar la amplitud de la salida de la neurona, y
4. Una *ganancia exterior* que determina el umbral de activación de la neurona.

2.2.3.3 Modelos de reconocimiento híbridos

Sabemos que el uso de HMM es una herramienta muy útil para modelar sistemas de reconocimiento de habla. Sin embargo han de tenerse en cuenta sus limitaciones, parcialmente paliadas por los *sistemas híbridos*. .

Vamos a explicar a continuación un sistema híbrido que combina HMM y RNA.

Las redes neuronales pueden diseñarse para realizar de forma muy satisfactoria numerosas tareas, clasificación, regresión, etc., e incluso pueden utilizarse en sistemas RAH sin estar asociadas a un HMM como hemos explicado anteriormente.

Bourlard et al [10]propusieron un modelo híbrido HMM/ANN para RAH continuo en el que se entrena un perceptrón multicapa MLP (del inglés "Multilayer Perceptron") para estimar las probabilidades a posteriori de los estados de un HMM M_i, dada un secuencia de observaciones X. Las probabilidades a posteriori pueden escribirse como:

$$\Pr(M_i|\, X) = \sum_{s_1^L} \Pr(s_1^L, M_i|X)$$

Donde s_1^L es la secuencia de estados recorridos. En la siguiente figura podemos ver la arquitectura básica de un modelo híbrido donde cada salida de la red estima la probabilidad a posteriori de cada estado:

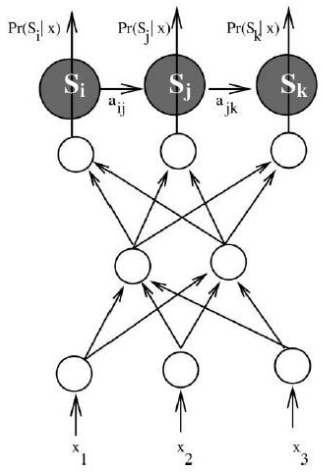

Figura 5: arquitectura básica de un modelo híbrido donde un ANN 2-capa "feedforward" estima las probabilidades a posteriori de los estados Si,Sj,Sk de izquierda-derecha HMM dado una hipotética observación acústica x=(x1,x2,x3).[11]

Distintos experimentos realizados con sistemas híbridos MLP/HMM señalan importantes mejoras respecto a los sistemas clásicos HMM[10]:

- Los sistemas híbridos relativamente sencillos han resultado muy eficientes (desde el punto de vista de consumo de CPU y requerimientos de tiempo de memoria), a la par que precisos.
- Entrenamos los sistemas más complejos de forma sencilla con muy buenos resultados en tareas complejas de reconocimiento continuo de voz.
- Utilizando un mismo número de parámetros se obtienen mejores resultados con un sistema híbrido que con un HMM convencional.
- La inclusión de información sobre el contexto en los sistemas MLP/HMM es sencilla y las componentes de los vectores de parámetros de entrada no tienen por qué ser estadísticamente independientes.

2.2.4 MODELADO DEL LENGUAJE

El denominado modelado del lenguaje es una parte esencial de los sistemas de RAH, puesto que es el encargado modelar la probabilidad de cada secuencia de palabras.

En el ámbito del reconocimiento de habla continuo será una pieza clave para que las palabras reconocidas tengan sentido en el contexto de una frase.

Existen distintos métodos de modelado de lenguaje, uno de los más sencillos y más utilizados son los basados en N-gramas, que recogen de forma sencilla, a partir de aproximaciones, las probabilidades de concatenaciones entre palabras. Así, el problema se reduce a calcular la probabilidad de las palabras en función de sus N-1 predecesoras.

2.3 DIFERENCIAS Y SIMILITUDES

Vistas las definiciones y procedimientos del reconocimiento máquina del habla tanto humano como podemos destacar entre otras las siguientes similitudes y diferencias entre ellos[12]. Podemos decir que el RHH y el RAH tienen en común los siguientes aspectos funcionales:

- Transforman la señal de audio de entrada en una representación sub-simbólica.
- Requieren un léxico o vocabulario.
- Requieren un mecanismo con el cual se puedan comparar los distintos modelos de palabras con la señal emitida.
- Requieren un proceso de búsqueda de palabras basado en la comparación de hipótesis.

Y a nivel conceptual:

- Los humanos y las máquinas realizan la misma tarea cuando decodifican una señal en una secuencia de elementos léxicos.

En cuanto a las diferencias cabe destacar:

- El reconocimiento humano de habla se basa en el entendimiento fundamental del habla humana mientras que el reconocimiento automático del habla se basa en la decodificación automática del habla humana mediante el uso de modelos estadísticos, de esta manera minimizamos las tasas de error de reconocimiento.

2.4 RECONOCEDOR DE HABLA USADO EN NUESTROS EXPERIMENTOS

Una vez introducidos los fundamentos básicos del reconocimiento automático del habla pasaremos a detallar el sistema de reconocimiento de habla que hemos usado en la realización de nuestro proyecto.

Su diagrama de bloques es el siguiente:

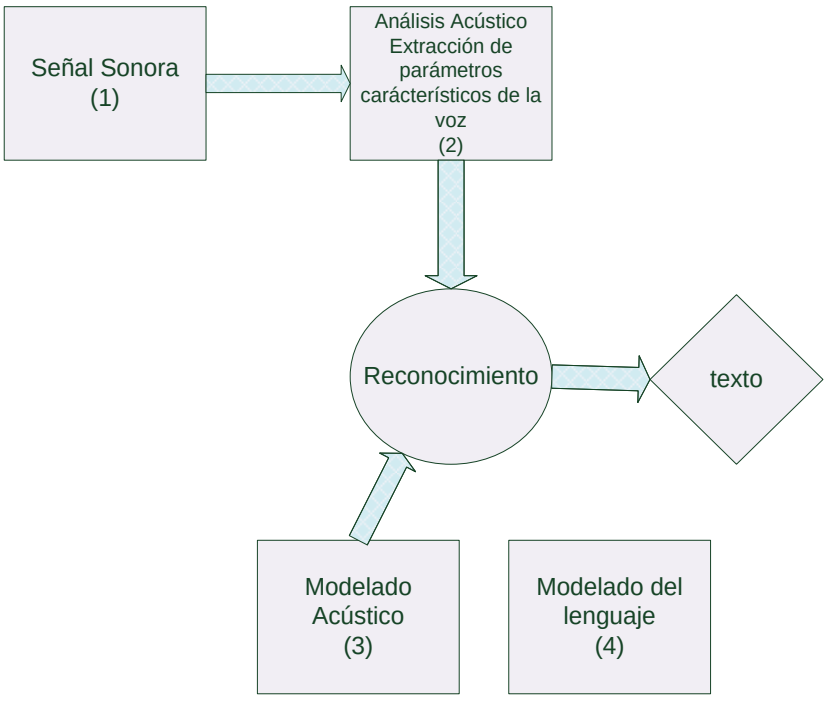

Figura 6: Diagrama bloques del RAH usado en nuestros experimentos

Donde:

(1) *Señal sonora*: Base de datos ISOLET. (Para más detalle véase apartado 4.1).

(2) *Extracción de parámetros característicos de la voz*: En nuestro caso obtendremos y trabajaremos con los coeficientes MFCC (en nuestro caso extraemos 27 características) y PLP (en nuestro caso extraemos 39 características) que son además los más comunes.

(3) *Modelado Acústico*: Utilizaremos un modelo híbrido en concreto HMM-MLP puesto que es el que mejor se adapta a nuestras necesidades

(4) *Modelado del lenguaje*: No tendremos un modelado de lenguaje como tal, ya que nuestro experimento se basa en reconocer palabras aisladas, por tanto no tendremos problemas de habla continua y no nos hará falta este bloque, de ahí que aparezca aislado en la figura.

3 HERRAMIENTAS PARA EL ANÁLISIS DE ERRORES.

Una vez que realizamos unos experimentos es muy importante poder analizar los resultados obtenidos para comprender el funcionamiento del sistema, sus debilidades y sus puntos fuertes.

Para ello tendremos que valernos de distintas herramientas que nos ayuden a analizar los resultados obtenidos por nuestro reconocedor y así poder sacar conclusiones sobre su funcionamiento y cómo mejorarlo. En este proyecto queremos ir más allá de la simple evaluación de la tasa de error del reconocedor que puede esconder comportamientos defectuosos de los procedimientos de aprendizaje[13],[14].

Para el análisis de los experimentos llevados a cabo en este estudio, pues, usaremos las herramientas detalladas a lo largo de este capítulo.

3.1 MATRICES DE CONFUSIÓN

La definición de una matriz de confusión es la siguiente: Sean

$$V_X = \{x_i\}_{i=1}^{n} \text{ y } V_Y = \{y_j\}_{j=1}^{p}$$

los conjuntos de etiquetas de los datos de entrada y salida, respectivamente, en una tarea de clasificación multiclase. El comportamiento del clasificador se puede mostrar a lo largo de N iteraciones del experimento para obtener la matriz $N_{XY.}$ Donde $N_{XY}(x_i, y_i) = N_{ij}$ representa el número de veces que ocurre el evento conjunto ($X = x_i., Y = y_i.$). Decimos que N_{XY} es la matriz de confusión o la tabla de contingencia de nuestro clasificador [15]

Cuando nos referimos a matrices de confusión en la temática del reconocimiento de habla podemos definirlas de la siguiente manera:

"Una matriz de confusión es una herramienta de visualización empleada para analizar las confusiones individuales entre las clases (ya sean unidades acústicas, palabras, etc.) en un reconocedor."

Gracias a esta herramienta podremos observar, por ejemplo, qué fonemas se confunden con otros y en qué medida lo hacen. Esta representación de los datos puede ayudarnos a averiguar cuantas veces reconocemos la locución emitida y cuantas veces la confundimos con cada una de las demás. En algunos casos dicha locución será una palabra, en otros una sílaba, un fonema o cualquier otra unidad acústica que queramos analizar.

En la matriz de confusión, las filas están indexadas en los símbolos de entrada y las columnas con los símbolos de salida. De esta forma podremos ver cuánto se confunden entre sí de forma pormenorizada con el conteo de la celda $N_{XY}(x_i, y_i) = N_{ij}$.

Existen distintas maneras de visualizar matrices de confusión.

Una de las representaciones más sencillas es la numérica, que podemos ver en la figura 7. Esta representación simplemente refleja el número de fonemas emitidos/recibidos correctamente y erróneamente.

	/p/	/t/	/k/	/f/	/th/	/s/	/sh/	/b/	/d/	/g/	/v/	/dh/	/z/	/zh/	/m/	/n/
/p/	72	68	90	20	15	4	1	2	4	1	0	1	0	0	0	2
/t/	73	72	74	20	8	6	3	1	2	2	0	2	0	1	0	0
/k/	63	74	127	9	7	5	2	0	0	1	0	1	1	1	0	1
/f/	7	7	10	63	69	41	8	3	1	1	1	3	0	1	1	0
/th/	5	8	11	60	85	45	14	2	4	2	6	5	1	0	0	0
/s/	1	6	5	19	49	125	60	5	2	1	2	9	4	0	0	0
/sh/	2	6	8	8	22	69	89	2	4	1	0	3	5	1	0	0
/b/	0	1	1	19	14	5	0	134	20	13	14	11	4	1	2	1
/d/	0	0	2	0	1	6	4	19	120	23	2	3	11	3	0	2
/g/	0	0	2	1	0	5	1	11	116	69	8	7	11	4	1	2
/v/	0	1	0	1	1	2	0	25	4	8	111	55	18	2	2	2
/dh/	0	1	1	6	5	1	0	43	16	15	75	66	23	11	1	4
/z/	2	0	2	1	5	5	2	21	20	17	18	33	91	25	1	1
/zh/	0	0	0	0	4	0	2	1	27	29	11	16	83	78	1	0
/m/	0	0	0	0	0	0	0	12	3	0	1	0	0	0	219	57
/n/	0	0	0	0	1	1	0	12	3	1	1	2	0	0	99	120

Figura 7: Matriz de confusión fonética para S/N=12db y frecuencia 200-400Hz M&N5

En nuestro estudio usaremos la representación que podemos observar en la figura 8, con un "mapa de calor" (Ing. "heatmap") en el que el nivel de gris representa el número de confusiones siendo el blanco el cero y el negro el máximo número de confusiones.

Como se puede observar en la figura 8, la matriz de confusión representada marca claramente la diagonal (que coincide con las veces que el reconocedor ha clasificado lo recibido como lo que realmente se ha emitido), con esta diagonal podemos sacar lo que se denomina *accuracy* (en castellano lo traduciremos como precisión), de este concepto hablaremos más detalladamente en este mismo apartado.

En la figura 8 (a)[2] vemos que existen grupos de confusión, es decir, hay fonemas que tienden a confundirse unos con otros por su similitud fonética. Para poder visualizarlos más cómodamente podemos ordenar la matriz de confusión como podemos ver en la figura 8(b). Este reordenamiento se ha llevado a cabo de forma semi-automatizada basándonos en la información de los retículos de confusión que presentaremos en la sección 3.3 pero Miller&Nicely hicieron este mismo trabajo de forma manual a base de observación, prueba y error, lo que les llevó a la definición de los cinco canales que presentamos en el apartado 2.1.3 de este documento.

[2] Hay que tener en cuenta que la tarea ISOLET consiste en el reconocimiento de las letras del alfabeto y aunque hay una relación con los fonemas que representan esta relación no es directa (véase la sección 4.1.2)

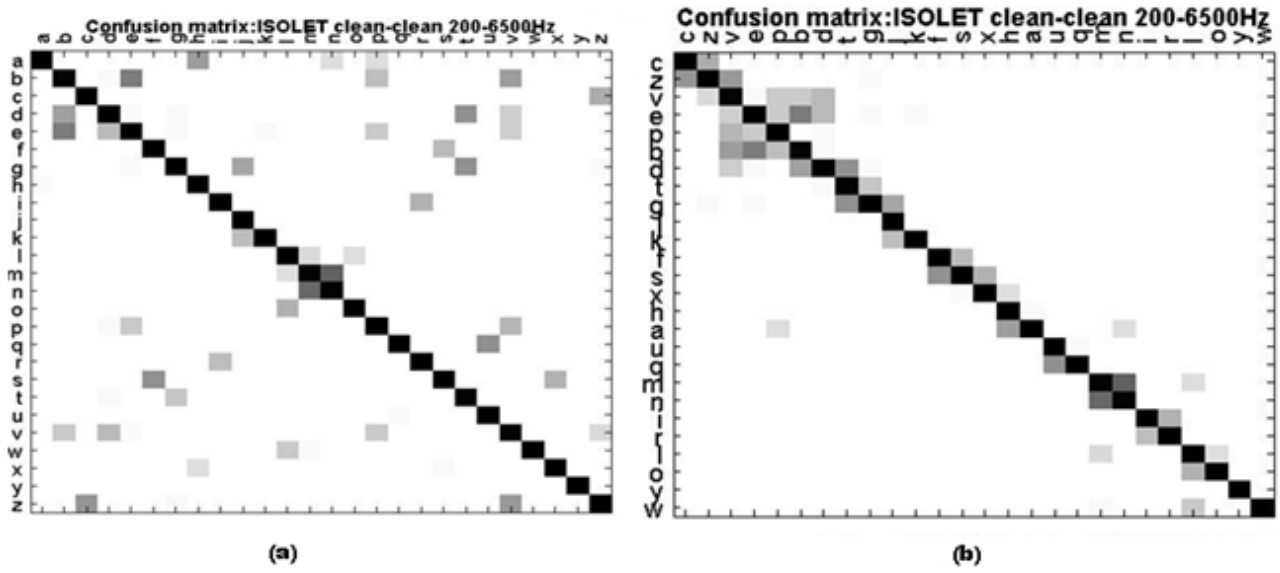

Figura 8: Representación matriz de confusión de la pronunciación de las letras del alfabeto en inglés (véase la descripción de ISOLET) sin ordenar (a) y ordenada (b).

En reconocimiento del habla existen distintos grupos de confusiones comunes, como por ejemplo el conocido grupo E-SET[16], la confusión entre la /m/ y la /n/ (ambas pueden apreciarse en las figuras 8 (a), (b)).

El llamado grupo E-SET[16] es el conjunto de 10 letras del alfabeto inglés que por su similitud fonética tienden a confundirse más fácilmente por los reconocedores de habla automática. Este es el conjunto de letras [b, c, d, e, g, p, t, v, z[3]]. En la figura 8 (b) se han ordenado los fonemas de tal forma que se ve claramente cuáles son las confusiones más comunes en el reconocedor. Podemos observar que la mayoría corresponden al grupo E-SET, dentro de este grupo podríamos agrupar las confusiones más comunes en parejas de fonemas como se ve en la figura, además de la pareja mencionada anteriormente /m/,/n/.

Estas parejas de fonemas serían las siguientes [16] /b/ y /d/; /p/ y /t/; /b/ y /e/, /m/ y /n/, /v/ y /z/ y por último /f/ y /s/.

3.1.1 CONCEPTO DE PRECISIÓN DEL RECONOCEDOR

Definimos la precisión en un reconocedor como la suma de todos los casos de acierto (la suma de la diagonal de la matriz de confusión resultante) entre el número total de entradas.

No debemos confundir la traducción de *accuracy* como precisión con el concepto de *precision* en inglés, normalmente utilizado en problemas de clasificación binaria o de detección.

Veamos la diferencia de accuracy y *precision* con un ejemplo [17].

Imaginemos que nos encontramos en un proceso de clasificación donde tenemos las siguientes posibilidades:

[3] La pronunciación en inglés americano de este fonema.

	actual class (observation)	
predicted class (expectation)	**tp** (true positive) Correct result	**fp** (false positive) Unexpected result
	fn (false negative) Missing result	**tn** (true negative) Correct absence of result

Figura 9: Tabla de probabilidades obtenida de [17]

Definimos precision (prec) y accuracy(acc) como:

$$prec = \frac{tp}{tp + fp}$$

$$acc = \frac{tp + tn}{tp + tn + fp + fn}$$

Como podemos ver la *precision* solo tiene en cuenta los casos positivos de acierto correctos entre los casos positivos totales.

Sin embargo la *accuracy* es la suma de todos los casos que se consideran correctos (tanto los positivos como los negativos) entre el número total de casos posibles.

Una vez aclarado esto nos referiremos como precisión teniendo en cuenta que nos referimos al concepto de *accuracy* y en todo caso, y puesto que nuestro problema es multiclase quedará definido como:

$$\hat{P}_{XY}(x, y) \approx \frac{N_{XY}(x, y)}{N}$$

En donde $\hat{P}_{XY}$ es la estimación de la probabilidad conjunta de X e Y, y N el número total de casos.

3.1.2 ¿ES LA PRECISIÓN UNA BUENA HERRAMIENTA DE MEDIDA PARA LA CALIDAD DE UN CLASIFICADOR?

El objetivo durante décadas en el diseño y desarrollo de los reconocedores de habla ha sido maximizar el valor de la precisión; es decir, cuanto mayor sea la precisión del reconocedor mejor será éste. Debemos tener en cuenta que al basarnos en el valor de la precisión como medida de calidad de un reconocedor solo tenemos en cuenta la diagonal de la matriz de confusión dejando de lado errores individuales que pueden ser importantes a la hora de evaluar nuestro clasificador.

Desde hace varios años son muchos los que se preguntan si la precisión en una buena medida de calidad en clasificadores. Según varios estudios [13],[14] ,[18] se ha demostrado que no por tener un valor de precisión mayor el clasificador es mejor, es decir, se demuestra que para ciertos conjuntos

de datos clasificadores con un valor de precisión menor obtienen mejores resultados que los que tienen un valor de precisión mayor.

En [13] se plantean las siguientes cuestiones ¿es la precisión la medida más adecuada para evaluar la calidad de un clasificador? Y si no lo es, ¿hemos estado fijándonos en la medida equivocada durante años?

En dicho artículo partiendo de las cuestiones mencionadas anteriormente, se sospecha que más de un tercio de los datos dados como aciertos para la obtención de la precisión se producen por casualidad, si esto es así, no por tener mayor precisión el reconocedor es mejor, puesto que puede estar dejando de lado características y comportamientos del clasificador más importantes a la hora de clasificar.

En el artículo [13] se comparan los resultados obtenidos basándose en la precisión del clasificador con los resultados obtenidos basándose en otra medida de calidad, que en este caso es la Kappa de Cohen[19] demostrándose que es mejor medida de calidad la Kappa de Cohen que la medida de la precisión, ya que tiene en cuenta los aciertos que tienen que ver con el clasificador dejando de lado los que pueden deberse a mera suerte; de esta forma la medida de la calidad del clasificador es más real.

Por tanto podemos decir que las cuestiones planteadas se corresponden con las sospechas del artículo concluyendo que evaluar la calidad de un reconocedor basándonos en la precisión no es la forma más óptima y que probablemente nos hemos equivocado en basar la calidad de un clasificador en la precisión durante todos estos años.

En esta línea y bajo las mismas cuestiones se hallan [14][18] en los que se nos introduce otra forma de medir la calidad para los clasificadores y que además ofrece una representación novedosa de la calidad del reconocedor sin basarse en la precisión. A esta nueva forma de representación ha sido bautizada como el triángulo entrópico (para más información véase el apartado EL TRIÁNGULO ENTRÓPICO Y OTRAS MEDIDAS DE PRESTACIONES de este documento) que será junto con los retículos de confusión, las herramientas que se usen en este proyecto para evaluar la calidad del clasificador/reconocedor.

En estos estudios también se demuestra que clasificadores con menor precisión clasifican mejor según qué conjunto de datos que los clasificadores que para ese mismo conjunto tienen un mayor nivel de precisión.

En la siguiente figura vemos un ejemplo de la representación novedosa de la precisión que nos introducen los artículos[14][18].

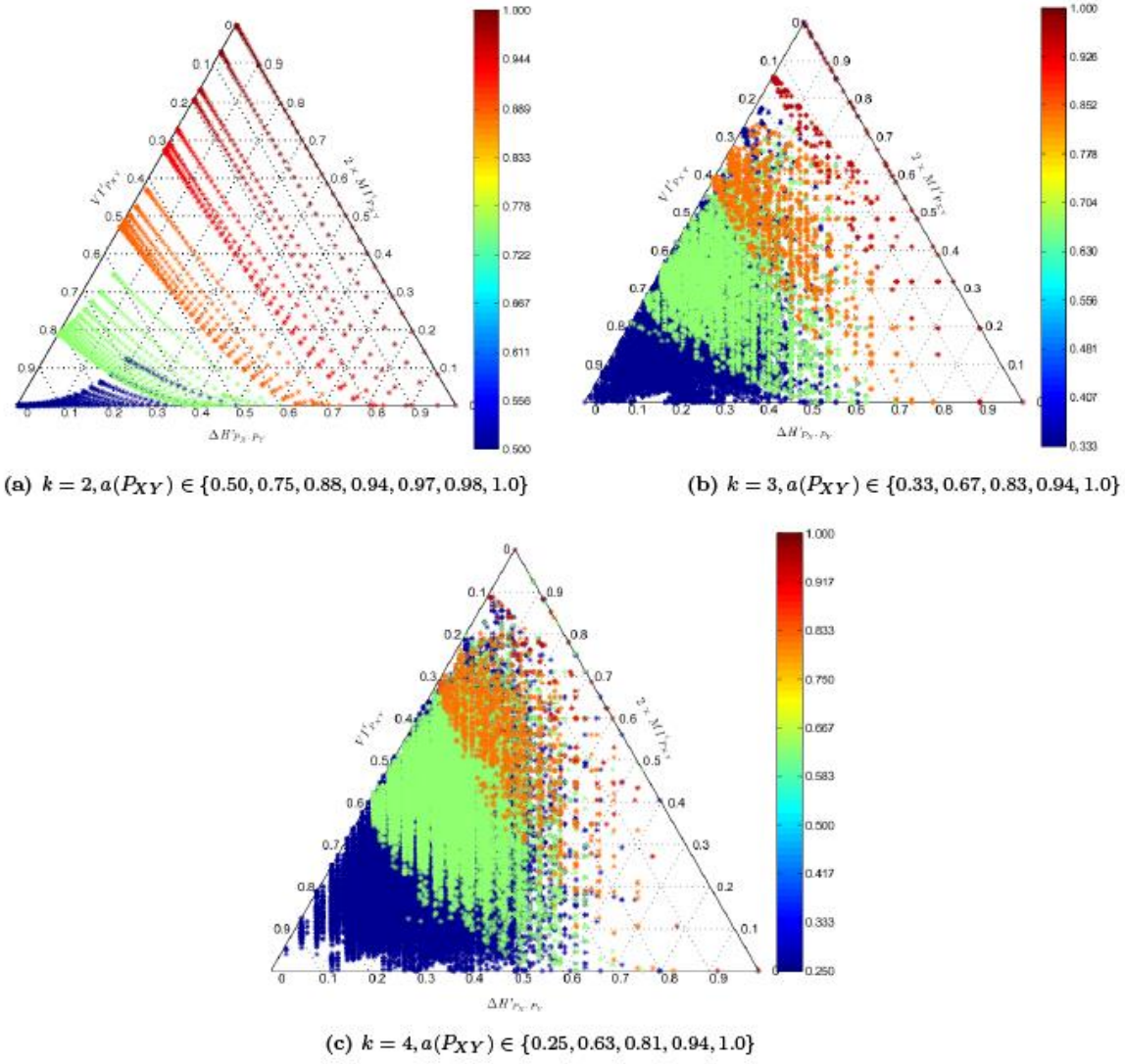

Figura 10: Representación Triángulos entrópicos obtenidas de [18]

En la figura vemos representada la descomposición de la entropía para dos matrices cuadradas de (a) k=2, (b) k=3 y (c) k=4. Representando las matrices de confusión para una tarea de clasificación a diferentes niveles de precisión que se ven descritos por la barra de color situada a la derecha de cada triángulo. La superposición de las áreas que representan a matrices con diferentes valores de precisión pero con valores de entropía similares es evidente a todos los niveles para k=3 y k=4 pero solo para valores bajos de precisión en el caso de k=2. Esto implica que la precisión no es un buen criterio para juzgar el flujo de información de las etiquetas de entrada para las etiquetas de salida de un clarificador. Podemos observar incluso que hay clasificadores que consiguiendo una precisión del 100% no transmiten información de la entrada a la salida.

De esta manera podemos empezar a pensar que realmente medir la calidad de los clasificadores basándonos únicamente en su precisión no es muy efectivo y puede resultar engañoso.

3.2 EL TRIÁNGULO ENTRÓPICO Y OTRAS MEDIDAS DE PRESTACIONES

El triángulo entrópico es una herramienta que analiza el comportamiento de los clasificadores multiclase (Ing. ”multiple-class o multi-class”) en términos de medidas de entropía de su matriz de confusión o tabla de contingencia[15][20].

- Primero: Se obtiene la ecuación de equilibrio de las entropías que contienen propiedades interesantes del clasificador.
- Segundo: Normalizando dicha ecuación obtendremos un simplex en un espacio entrópico tridimensional y después el diagrama entrópico de Finetti o triángulo entrópico.

Pasamos ahora a desarrollar lo introducido anteriormente.

El triángulo entrópico usa la descomposición de la entropía conjunta de dos variables aleatorias.

Podemos estudiar el rendimiento de nuestras matrices de confusión basándonos en estimaciones empíricas de la distribución conjunta entre los elementos de entrada y salida del clasificador, como la estimación de máxima verosimilitud empleada en la sección anterior:

$$\hat{P}_{XY}(x,y) \approx \frac{N_{XY}(x,y)}{N}$$

Donde:

- $\mathrm{P_{XY}}(\mathrm{x},\mathrm{y})$ es una estimación de la función de distribución entre las variables de entrada y salida con probabilidades marginales:$P_x(x) = \sum_{y_i \ni Y} P_{X,Y}(x, y_i)$ y $P_Y(Y) = \sum_{x_i \ni X} P_{X,Y}(x_i, y)$.
- $Q_{XY} = P_X . P_Y$ es la función de distribución con las mismas probabilidades marginales, $\mathrm{P_{XY}}$ considerándolas variables independientes.
- $U_{XY} = U_X . U_Y$ es el producto de las funciones de distribución uniformes (y por tanto, máximamente entrópicas) de X e Y, $U_x(x) = 1/n$ y $U_Y(y) = 1/p$, donde n y p son el número de clases de entrada y salida respectivamente.

Entonces la disminución de incertidumbre desde $\mathrm{U_{XY}}$ a $\mathrm{Q_{XY}}$ es la diferencia en entropías:

$$\Delta H_{P_X . P_Y} = H_{U_X . U_Y} - H_{P_X . P_Y} \quad (1)$$

Intuitivamente, $\Delta H_{P_X . P_Y}$ mide cuán lejos está el clasificador operando de la situación más general posible donde todas las entradas son equiprobables, lo que impide al clasificador especializarse en una clase sobrerrepresentada en detrimento de la precisión de clasificación en otras. Dado que $H_{U_X} = \log n$ y $H_{U_Y} = \log p$, $\Delta H_{P_X . P_Y}$ puede variar desde un mínimo $\Delta H^{min}_{P_x . P_y} = 0$, cuando las probabilidades marginales son uniformes $P_X = U_X$ y $P_Y = U_Y$, a un valor máximo $\Delta H^{max}_{P_x . P_y} = \log n + \log p$, cuando son distribuciones delta de Kronecker.

Queremos relacionar este decremento de la entropía con la información mutua $MI_{P_{XY}}$ esperada de una distribución conjunta. Con este propósito, nos damos cuenta que la formula de la información mutua:

$$MI_{P_{XY}} = \sum_{x,y} P_{XY}(x,y) log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \quad (2)$$

describe la disminución en la entropía resultante de pasar del producto de dos distribuciones independientes $Q_{XY} = P_X . P_Y$ a la distribución conjunta P_{XY},

$$MI_{P_{XY}} = H_{P_X . P_Y} - H_{P_{XY}} \ (3)$$

Y finalmente invocamos la conocida formula que relaciona la entropía conjunta $H_{P_{XY}}$ y la información mutua $MI_{P_{XY}}$ esperada con las entropías condicionadas de X dada Y, $H_{P_{X|Y}}$, y de Y dada X, $H_{P_{Y|X}}$.

$$H_{P_{XY}} = H_{P_{X|Y}} + H_{P_{Y|X}} + MI_{P_{XY}} \ (4)$$

Por lo tanto $MI_{P_{XY}}$ irá desde $MI_{P_{XY}}^{min} = 0$ cuando $P_{XY} = P_X . P_Y$, un mal clasificador, a un máximo teórico $MI_{P_{XY}}^{max} = (\log n + \log p)/2$ en caso de que las probabilidades marginales sean uniformes y que las variables de entrada y salida sean totalmente dependientes, un clasificador excelente.

La variación de información se define en [20] como

$$VI_{P_{XY}} = H_{P_{X|Y}} + H_{P_{Y|X}} \ (5)$$

Para clasificadores óptimos con relación determinista entre las variables de entrada y salida, y una matriz de confusión diagonal, $VI_{P_{XY}}^{min} = 0$.Al contrario, cuando son independientes $VI_{P_{XY}}^{max} = H_{P_X} + H_{P_Y}$.

Mezclando las ecuaciones (1)-(5) obtenemos la ecuación de equilibrio para la información perteneciente a una distribución conjunta

$$H_{U_{XY}} = \Delta H_{P_X . P_Y} + 2MI_{P_{XY}} + VI_{P_{XY}} \ (6)$$

La ecuación de equilibrio sugiere un diagrama de información como el que esta explicado en la figura 11.

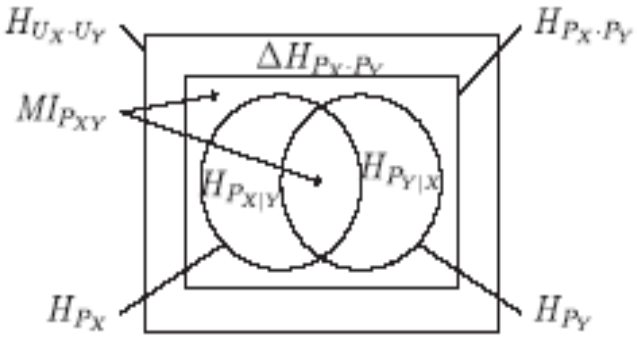

Figura 11: Diagrama de información extendido de las entropías correspondientes a una distribución bivariante. [20]

En este diagrama podemos identificar la conocida descomposición de la entropía conjunta $H_{P_{XY}}$ como las dos entropías H_{P_X} y H_{P_Y} cuya intersección es $MI_{P_{XY}}$.Nos damos cuenta que el incremento entre $H_{P_{XY}}$ y $H_{P_{X \cdot P_Y}}$ es otra vez $MI_{P_{XY}}$, por consiguiente la información mutua esperada aparece dos veces en el diagrama. Además, el interior del rectángulo exterior representa $H_{U_{X \cdot U_Y}}$, el interior del rectángulo interior $H_{P_{X \cdot P_Y}}$ y finalmente $\Delta H_{P_{X \cdot P_Y}}$, representa la diferencia de áreas entre los dos rectángulos.

Desarrollando la idea de la descomposición de la entropía sugerida por la ecuación de equilibrio, de la eq (6) y los párrafos siguientes a (1)-(5) obtenemos:

$$H_{U_{XY}} = \Delta H_{P_{X \cdot P_Y}} + 2MI_{P_{XY}} + VI_{P_{XY}}$$

$$0 \leq \Delta H_{P_{X \cdot P_Y}}, 2MI_{P_{XY}}, VI_{P_{XY}} \leq H_{U_{XY}}$$

que impone limitaciones estrictas en los valores que las cantidades pueden tomar, la más sobresaliente que dadas dos cantidades, la tercera viene fijada. Así, normalizando $H_{U_{XY}}$ obtenemos

$$1 = \Delta H'_{P_{X \cdot P_Y}} + 2MI'_{P_{XY}} + VI'_{P_{XY}} \quad (7)$$

$$0 \ll \Delta H'_{P_{X \cdot P_Y}}, 2MI'_{P_{XY}}, VI'_{P_{XY}} \leq 1$$

Esto es un simplex en el espacio normalizado $\Delta H'_{P_{X \cdot P_Y}} \times 2MI'_{P_{XY}} \times VI'_{P_{XY}}$. El diagrama de Finetti es una proyección de este simplex a 2 dimensiones como aparece en la figura 12, por lo que cada clasificador con distribución conjunta P_{XY} se puede caracterizar por sus *fracciones de entropía conjunta*, $F_{XY}(P_{XY}) = \left[\Delta H'_{P_{X \cdot P_Y}}, 2MI'_{P_{XY}}, VI'_{P_{XY}}\right]$

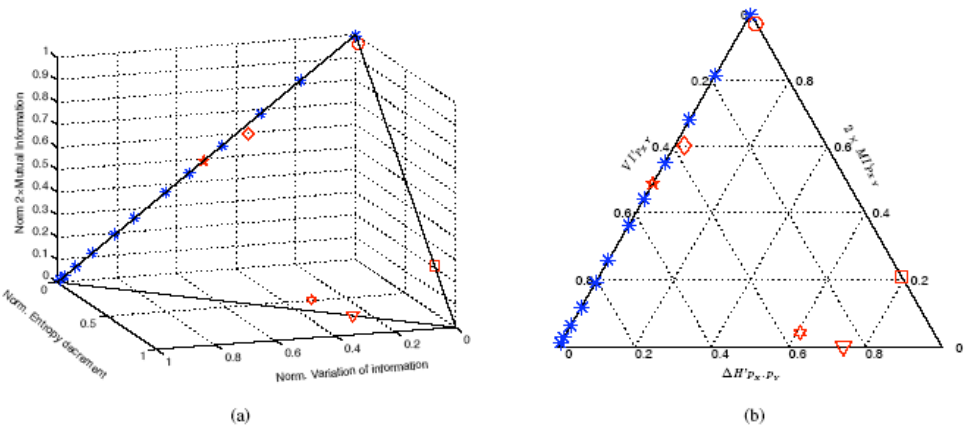

Figura 12: Representaciones entrópicas para distribuciones bivariables. (a) 2-simplex tridimensional; (b) diagrama entrópico de Finetti o triángulo entrópico[20]

3.2.1 EL TRIÁNGULO ENTRÓPICO SEPARADO

Observando en la ecuación (7), dado que tanto $\mathbf{U_X}$, $\mathbf{U_Y}$, $\mathbf{P_X}$, $\mathbf{P_Y}$ son independientes y las probabilidades marginales de $\mathbf{U_X}.\mathbf{U_Y}$ y $\mathbf{P_X}.\mathbf{P_Y}$ también, podemos escribir:

$$\Delta H_{P_X \,.\, P_Y} = \left(H_{U_X} - H_{P_X}\right) + \left(H_{U_Y} - H_{P_Y}\right)$$
$$= \Delta H_{P_X} + \Delta H_{P_Y},$$

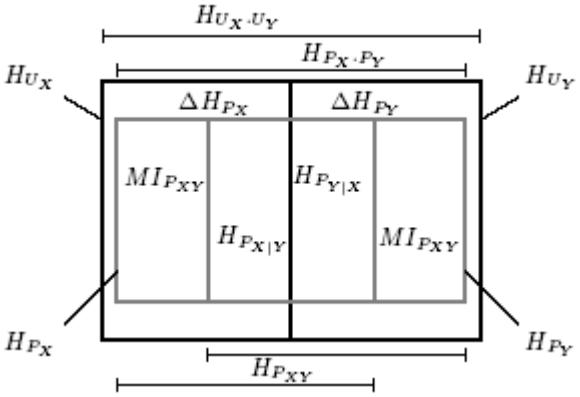

Figura 13: Representación de probabilidades de entropía separadas[20]

Lo que nos sugiere que podemos escribir por separado la ecuación de equilibrio (figura 13) para cada variable.

$$H_{U_X} = \Delta H_{P_X} + MI_{P_{XY}} + H_{P_{X|Y}}$$

$$H_{U_Y} = \Delta H_{P_Y} + MI_{P_{XY}} + H_{P_{Y|X}}$$

De esta manera podemos representar los datos de entrada y salida en el mismo triángulo entrópico pero separados. Lo que nos proporciona una herramienta valiosa a la hora de analizar las prestaciones de un reconocedor.

En nuestro proyecto para el análisis de los experimentos usaremos los triángulos entrópicos separados o Triángulos entrópicos divididos.

Para la comprensión de los mismos vamos a mostrar a continuación un ejemplo de análisis de un triángulo dividido.

3.2.1.1 EJEMPLO DE ANÁLISIS DE TRIÁNGULO ENTRÓPICO SEPARADO.

Lo primero que debemos saber es lo que estamos representando y cómo:

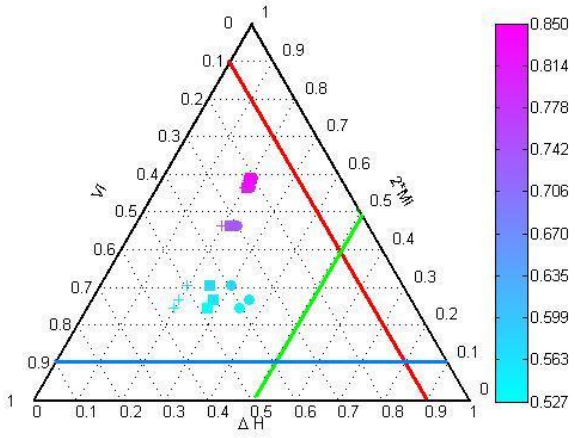

Figura 14: Ejemplo Triángulo entrópico dividido para MFFCC_clean paso bajo fonemas completos

Lo primero que debemos de tener en cuenta que el triángulo se forma basándonos en tres magnitudes de información que son las que podemos ver indicadas en los lados correspondientes del triángulo

- VI: variación de información. (Se lee su valor en la matriz sobre las líneas paralelas a la línea roja y con la escala de la izquierda)
- ΔH: incremento entropía .(Se lee su valor en la matriz sobre las líneas paralelas a la línea verde y con la escala de abajo)
- MI: información mutua. (Se lee su valor en la matriz sobre las líneas paralelas a la línea azul y con la escala de la derecha)

Además nos damos cuenta que se asocia a cada entrada y salida un grado de color que indica el nivel de precisión del reconocedor para esos datos de entrada y salida y que se puede interpretar con la ayuda de la barra de color que aparece a la derecha.

La presentamos de esta forma complementaria porque debemos tener en cuenta, como hemos dicho anteriormente, que de las magnitudes del triángulo no se puede deducir la precisión del reconocedor y que de esta representación no puede relacionarse ni obtenerse su valor.

En el caso representado (3.2.1) tenemos que:

- Las cruces representan las variables a la entrada (VI_X, ΔH_X).
- Los círculos representan las variables a la salida del reconocedor (VI_Y, ΔH_Y).
- Los cuadrados representan las variables conjuntas de entrada y salida y siempre se encuentra en el punto intermedio entre los datos de entrada y salida.

Al representarlos en el mismo triángulo y de forma separada podemos observar y analizar el comportamiento del reconocedor viendo cómo eran los datos de entrada y donde se sitúan a la salida del mismo.

Para saber si nuestro reconocedor es un buen clasificador y aprende, o por el contrario no lo es, debemos saber dónde se deben encontrar los datos para que el triángulo lo considere buen clasificador o no. Para explicar esto nos ayudaremos de la siguiente figura:

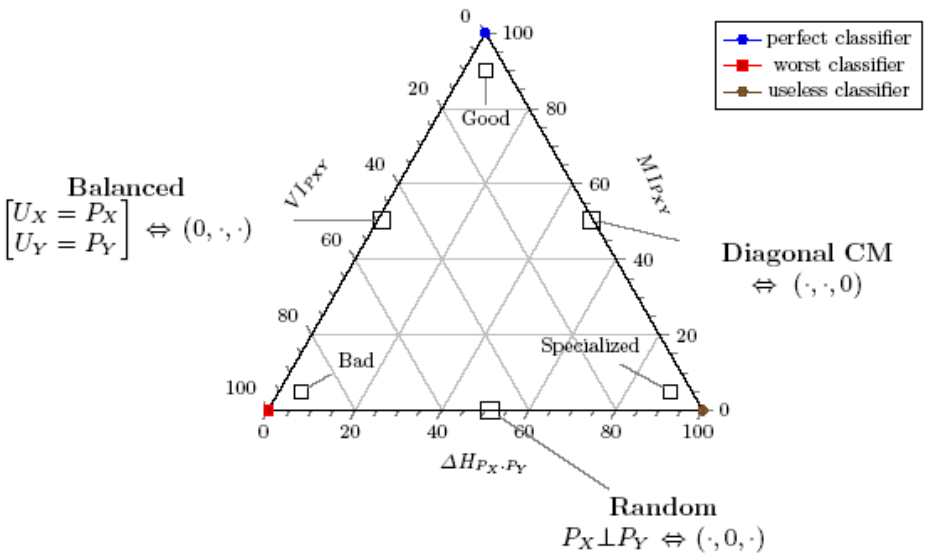

Figura 15: representación esquemática del ET donde se muestran las zonas interpretables y los casos extremos del mismo

Como podemos observar en la figura el mejor reconocedor será aquel que este en el vértice más alto del triángulo habiendo llevado una progresión lo más cercana a la izquierda posible, es a la izquierda del triángulo donde las condiciones de reconocimiento son más difíciles, por tanto si los resultados del reconocedor se encuentran pegados a la izquierda podremos decir que este se comporta de forma óptima en los casos más complicados.

En contraposición el peor reconocedor es aquel que se encuentre en la zona baja del triángulo lo más pegado a la izquierda, además decimos que todos aquellos resultados que se encuentren en la zona central-baja del triángulo son aleatorios, es decir, el reconocedor decide los fonemas reconocidos de forma aleatoria.

Además aquellos resultados que se acerquen a la derecha del triángulo lo harán buscando la especialización, suele buscarse la especialización agrupando los fonemas que no son fácilmente reconocibles a las clases mayoritarias obteniendo así mejores resultados.

3.3 ANÁLISIS FORMAL DE CONCEPTOS GENERALIZADO

Durante décadas, las matrices de confusión se han utilizado como herramienta de análisis para el reconocimiento de habla humano. Sin embargo, en pocas ocasiones se utilizan para el reconocimiento automático del habla debido a la falta de un procedimiento sistemático para su exploración.

En [21] se presenta una herramienta denominada Análisis de Conceptos Formales (Ing., "Formal Concept Analysis", FCA) generalizado que transforma las matrices de confusión en retículos de eventos de confusión ordenados, y que ha permitido verificar muchos de los resultados clásicos en RHH que identifican una jerarquía de los canales articulatorio-acústicos virtuales. Traduciendo esta técnica al RAH, se puede trazar un mapa detallado de las relaciones a través de las unidades del habla empleadas en el sistema para hacer que las diferentes fuentes de confusión se manifiesten de forma automatizada: influencia del léxico, errores de segmentación, variaciones de dialecto o limitaciones en la extracción de características, entre otras.

Para construir los retículos que hemos mencionado anteriormente; debemos seguir los siguientes pasos que explicaremos con ayuda de un ejemplo práctico (tomado de [21]).

1) Primeramente deberemos construir lo que denominamos matriz booleana de confusión (para entrar en más detalle sobre como hallar dicha matriz véase sección IV de [21]). Este proceso depende de un parámetro φ que es necesario explorar. A partir de ahora nos referiremos a ella como I_{CM}.

2) Para una matriz booleana de confusión I_{CM}, la tupla (V_A, V_B, I_{CM}) es lo que denominamos un *contexto formal*, y asumimos que éste codifica toda la información pertinente al fenómeno que está siendo analizado.

Tomaremos como ejemplo para nuestra explicación la matriz de confusión de la figura 16, extraída de los experimentos realizados por Miller y Nicely (1955)[1] tal y como se explica en[21].

	/p/	/m/	/t/	/f/	/th/	/k/	/s/
/p/	×		×			×	
/m/		×					
/t/	×		×			×	
/f/				×	×		
/th/				×	×		
/k/	×		×			×	
/s/					×		×

Figura 16: Matriz de confusión

El contexto formal como hemos definido anteriormente está formado por:

- V_A, que corresponden con los estímulos, los denominaremos a partir de ahora "objetos" para seguir la notación habitual en FCA.
- V_B, que corresponden con las respuestas, los denominaremos a partir de ahora "atributos".
- I_{CM}conocida como matriz de confusión booleana pasará a denominarse incidencias.

3) Denominaremos *conceptos formales* a parejas de un conjunto particular de estímulos que se confunden con un conjunto particular de respuestas, y viceversa.

Por ejemplo, podemos sacar los siguientes *conceptos* de la figura 16 utilizando algoritmos estándar de FCA[22]

- C1= ({**/s/**}, {/s/, /th/})
- C2= ({/**s**/, /**f**/, /**th**/}, {/th/})

Para distinguir entre estímulos y respuestas, pondremos los estímulos en negrita.

Al conjunto de estímulos de un concepto se le denomina extensión (en inglés, "extent") y al de respuestas intensión (en inglés, "intent").

4) El teorema básico del Análisis de Conceptos Formales afirma que un conjunto de conceptos formales de un contexto formal es un retículo completo llamado *retículo conceptual* $\mathcal{B}$ (V_A, V_B, M). Podemos referirnos al retículo conceptual de una matriz de confusión como su retículo de confusión.

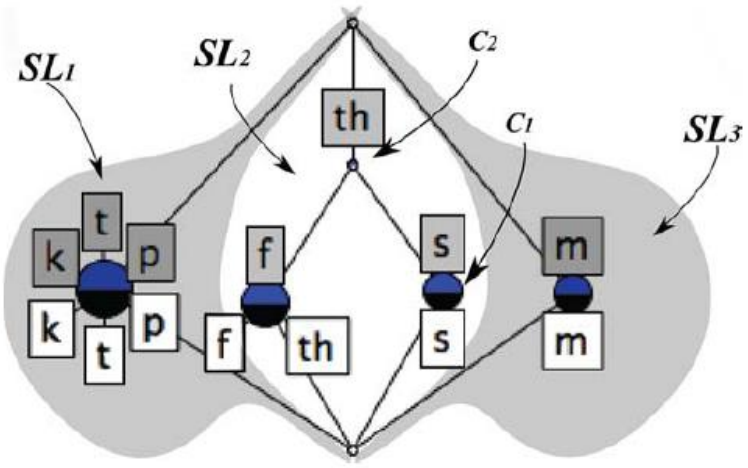

Figura 17: Representación reducida del retículo obtenida de figura 16

Representaremos los retículos de confusión mediante el uso de diagramas de Hasse [23] que fueron desarrollados para describir fácilmente órdenes parciales.

Los retículos de confusión se representan de la siguiente manera:

- Los nodos en el diagrama representan los *conceptos*.
- Los enlaces entre ellos, representan el orden parcial jerárquico entre sus vecinos adyacentes.

- El diagrama de Hasse de un retículo de confusión debe leerse en el eje vertical para descubrir las relaciones de confusión entre los estímulos en *las extensiones* y las respuestas en las intensiones.
- Las etiquetas de los estímulos se representan mediante cajas blancas, justo debajo del concepto correspondiente, y las etiquetas de respuestas son representadas mediante cajas grises justo encima. Para disminuir la confusión en el diagrama, en vez de etiquetar completamente cada nodo lo que haremos será poner la etiqueta de cada respuesta solo en el concepto más alto que aparezca (más abstracto) y la etiqueta de cada estimulo sólo en el más bajo (más específico). A esto lo denominamos *representación reducida* (de etiquetas), como la mostrada en la figura 17.

Para entender mejor dicha representación y las confusiones que se presentan en la misma debemos tener en cuenta los distintos tipos de fonemas, como se clasifican según su fonética y los grupos de confusiones más comunes entre ellos mismos.

Para poder comprender mejor los retículos de confusión nos ayudaremos de la siguiente figura en la que están representados los fonemas del experimento de M&N55 desde el punto de vista del modo (filas) y punto (columnas) de articulación.

De la figura 18 podemos observar como el retículo de la Figura 17 refleja el canal del modo de articulación claramente separando en distintos sub-retículos adjuntos[4] (en la figura representados como SL_1, SL_2 y SL_3) los fonemas oclusivos, los fricativos y los nasales, reflejando que dentro de cada uno de los grupos las confusiones son más numerosas que entre los grupos.

	labial	lab-dental	dental	alveolar	pal-alv	velar
plosive	/p/ /b/			/t/ /d/		/k/ /g/
fricative		/f/ /v/	/th/ /dh/	/s/ /z/	/sh/ /zh/	
nasal	/m/			/n/		

Figura 18: Representación de las consonantes inglesas: tipo de consonante según el modo y el punto de articulación al pronunciar los fonemas.

Para nuestro análisis tendremos también en cuenta a las confusiones más comunes enumeradas en el 3.1 de este documento: conjunto E-SET y parejas de consonantes más comunes a la hora de confundirse entre sí.

Cuando estudiemos los retículos de confusión obtenidos en la realización de nuestros experimentos tendremos estas referencias en cuenta para el estudio de las distintas confusiones de nuestro clasificador y el comportamiento de las mismas ¿se comporta de forma acorde a la teoría fonética o hay otros factores que influyen en la confusión?

Gracias a esto, analizaremos los retículos de confusión de nuestro experimento obteniendo así conclusiones sobre cómo se confunde el clasificador e información de interés para mejorar los clasificadores en un futuro.

[4] Llamamos sub-retículos adjuntos a aquellos que sólo comparten el *top* (supremo) y el *bottom* (ínfimo) del retículo.

4 EXPERIMENTOS Y RESULTADOS

4.1 DESCRIPCIÓN DE LOS EXPERIMENTOS.

4.1.1 DESCRIPCIÓN DEL EXPERIMENTO DE RHH

En este apartado pasaremos a explicar con detalle los experimentos realizados por Miller & Nicely.

Como ya hemos dicho en el apartado 2 de este documento para el RHH nuestra referencia son los experimentos llevados a cabo por Miller & Nicely[1].

El conjunto de fonemas escogidos para la elaboración de los experimentos es el siguiente:

/p /, /t /, /k /, /f/, /θ /, /s/, /ʃ/, /b/, /d /, /g /, /v /, /ð /, /z /, /ʒ /, /m/ y /n/

Fue elegido en su momento después de observar que eran los que más confusiones generaban en el reconocimiento y su estudio resultaba más interesante para sacar nuevas conclusiones ya que la información existente sobre el reconocimiento del conjunto de las vocales era más extensa.

Los sujetos elegidos para llevar a cabo los experimentos poseían las siguientes características:

Cinco mujeres como hablantes y oyentes, de nacionalidad norteamericana excepto una canadiense, sin defectos de audición o dicción y capaces de pronunciar las dieciséis sílabas requeridas sin sentido sin acento notable.

Puesto que el experimento se prolongó durante varios meses algunos sujetos de estudio fueron reemplazados por otros (se tuvo cuidado en aleccionar a los nuevos sujetos de la mecánica del experimento para que no ralentizaran ni enturbiasen el mismo). La mecánica del experimento es sencilla, un sujeto pronuncia distintas listas, previamente elaboradas de 200 sílabas sin sentido de tal manera que la probabilidad de cada sílaba sea de 1 en 16, mientras el resto escucha e interpreta lo que recibe. Se eligen listas aleatorias en cada una de las repeticiones del experimento.

Las consonantes inicialmente se emitían delante de la vocal /a/ (como en 'father').

Las silabas duraban en promedio 2.1 segundos y se obligaba a los oyentes a responder (adivinar si fuese necesario) para cada silaba escuchada.

Con 4 oyentes se obtienen 800 eventos sílaba-respuesta por emisor para estudiar las confusiones.

Juntando los cinco sujetos de estudio se dispone de un total de 4000 observaciones para cada condición probada.

Para cada receptor se genera una matriz de confusión donde se ve qué silaba ha sido pronunciada/emitida y cual se ha dado como respuesta. Cada celda de la matriz, por lo tanto, contiene el número de veces (conteos) que se ha proporcionado una respuesta a un determinado estímulo para cada una de las 16x16=256 posibles parejas de estímulo-respuesta.

Las grabaciones se llevaron a cabo usando un micrófono WE-633A, preservando siempre la misma distancia al micrófono de los distintos sujetos del experimento.

Dichas grabaciones fueron amplificadas, filtradas (a las frecuencias de interés), mezcladas con ruido, amplificadas otra vez y presentadas a los receptores mediante el uso de unos auriculares PDR-8.

Se realizaron dos tipos de experimento: añadiendo ruido y eliminando parte del contenido frecuencial mediante filtrado.

En el primer caso, los experimentos se replicaron estudiando distintas SNR, para ver el comportamiento de los sujetos respecto al ruido. Las SNR estudiadas fueron -18db, -12db, -6db, 0db, 6db y 12db.

En el segundo caso, quisieron averiguar el comportamiento de reconocimiento modificando las frecuencias de corte superior e inferior en las señales de voz grabadas:

- Se fijó una frecuencia de corte inferior de 200 Hz variando la frecuencia de corte superior (300, 400, 600, 1200, 2500, 5000, 6500) Hz.
- Así mismo se fijó una frecuencia de corte superior de 5000 Hz variando la frecuencia de corte inferior (1000, 2000, 2500, 3000, 4500) Hz.

La SNR para este análisis se fijó en 12db (cfr. [1] sección "results").

De este modo obtuvieron un total de 17 matrices de confusión con las que más tarde realizaron sus estudios y sacaron las conclusiones resumidas en el apartado 2.1. En nuestro estudio también utilizaremos las matrices de confusión obtenidas en estos experimentos para realizar un estudio comparativo entre éstos y los obtenidos con los experimentos realizados por nosotros (detallados en el aparatado 4.1.2 de este documento).

4.1.2 DESCRIPCIÓN DEL EXPERIMENTO RAH

Uno de los objetivos de este proyecto es analizar el comportamiento de un reconocedor automático del habla híbrido replicando el análisis frecuencial llevado a cabo en los experimentos realizados M&N55 [1] comparándolo a su vez con los resultados obtenidos por éstos.

Para ello se llevaron a cabo los experimentos descritos en este apartado.

En primer lugar elegimos la base de datos ISOLET (para más información léase apartado 4.1.2.1 de este documento).

Seguidamente, de forma análoga a nuestro experimento de referencia, modificamos las señales de voz de nuestra base de datos mediante filtrados en frecuencia. De forma idéntica a la de M&N55 se ha filtrado la base de datos con las siguientes frecuencias:

- Se fijó una frecuencia de corte inferior de 200Hz variando la frecuencia de corte superior (300, 400, 600, 1200, 2500, 5000, 6500) Hz.
- Así mismo se fijó una frecuencia de corte superior de 5000 Hz variando la frecuencia de corte inferior (1000, 2000, 2500, 3000, 4500) Hz.

Una vez obtenida la base de datos filtrada a las frecuencias mencionadas anteriormente procedimos a la realización de los siguientes pasos:

PASO 1. Instalación en los ordenadores donde se llevaron a cabo los experimentos el paquete Sprachcore_nogui (para más detalle véase apartado 4.1.2.2 de este documento) donde se encuentra el reconocedor utilizado y los distintos programas en los que nos hemos basado para la realización de estos experimentos.

PASO 2. Mediantes las herramientas Feacat y Feacalc obtuvimos los coeficientes MFCC y PLP de cada base de datos filtrada y los preparamos para poder analizarlos en formato pfile.

PASO 3. Una vez hecho esto normalizamos los datos obtenidos mediante la herramienta norms.

PASO 4. Una vez realizados los pasos anteriores procedimos a usar el reconocedor para obtener nuestros resultados.

Usamos el reconocedor de habla automático híbrido basado en [10]. En esta aproximación los MLPs (multi-layer perceptrons) se usan para el modelado acústico.

Para la validación de los resultados y dado que la base de datos es bastante reducida utilizamos una estrategia de validación cruzada dividiendo el conjunto total en 5 subconjuntos (eng. *folds*). Así, realizamos 5 entrenamientos independientes donde 4 de los subconjuntos conforman el conjunto de entrenamiento y el quinto, el de test. Dado que además es necesario ajustar una serie de parámetros para la decodificación de Viterbi utilizamos el primero de los experimentos (test en *fold*1, entrenamiento con *folds* 2-5) como conjunto de desarrollo que, por tanto, debe eliminarse del resultado final. Así, los resultados presentados serán siempre el promedio de los test realizados en los subconjuntos 2-5 (*folds* 2-5).

PASO 5. Obtuvimos las matrices de confusión correspondientes a los fonemas para cada tipo de coeficiente (MFCC, PLP) a nivel de trama (es decir, a la salida de las redes neuronales) mediante el uso de un script que invoca al reconocedor anteriormente mencionado.

PASO 6. Obtuvimos las matrices de confusión a nivel de palabra (en este caso correspondientes a las letras del alfabeto) de la base de datos para cada tipo de coeficientes (MFCC, PLP) mediante un script que invoca al reconocedor anteriormente mencionado.

PASO 7. Preparamos de forma adecuada las matrices de confusión obtenidas para poder analizar los resultados mediante la herramienta MATLAB.

En este punto decidimos analizar exclusivamente los resultados de los experimentos correspondientes a los fonemas (a nivel de trama) dejando los otros para futuros estudios.

Dado que tenemos interés en comparar estos resultados con los obtenidos en M&N55 obtenemos distintas submatrices de las matrices de confusión originales:

1) ISOLET COMPLETO: Conjunto isolet sin reducir de dimensión 27x27

/aa/, /ae/, /ah/, /ax/, /ay/, /b/, /ch/, /d/, /eh/, /ey/, /f/, /iy/, /jh/, /k/, /l/, /m/, /n/, /ow/, /p/, /r/, /s/, /t/, /uw/, /v/, /w/, /y/, /z/

2) ISOLET-M&N55: Subconjunto de alófonos comunes entre ISOLET y los experimentos de M&N55 de dimensión 11x11.

/b/, /d/, /f/, /k/, /m/, /n/, /p/, /s/, /t/, /v/, /z/

3) ISOLET VOCALES: Subconjunto de las vocales de dimensión 10x10.

/aa/, /ae/, /ah/, /ax/, /ay/, /eh/, /ey/, /iy/, /ow/, /uw/

4) ISOLET CONSONATES: Subconjunto de las consonantes de la base de datos isolet de dimensión 17x17.

/b/, /ch/, /d/, /f/, /jh/, /k/, /l/, /m/, /n/, /p/, /r/, /s/, /t/, /v/, /w/, /y/, /z/

Estas reducciones de los datos se han hecho una vez realizado el experimento debido a la complejidad de nuestro entorno de experimentación; explicaremos lo que esto implica en la sección 4.1.3.

PASO 8. Escogimos para el análisis los resultados obtenidos de los experimentos basados en los coeficientes MFCC, de estos resultados obtuvimos *triángulos entrópicos* y una vez obtenidos y analizados pasamos a elaborar, para su posterior análisis, los *retículos de confusión* para una selección de las frecuencias más interesantes para nuestro análisis.

PASO 9. Una vez obtenidos tanto los retículos como los triángulos entrópicos procedimos a analizar los resultados y sacar conclusiones.

4.1.2.1 DESCRIPCIÓN DE LA BASE DE DATOS

La base de datos utilizada para la realización de los experimentos de reconocimiento automático de habla (RAH) es la base de datos conocida como ISOLET [24] modificada.

ISOLET es una base de datos que contiene 7800 realizaciones de las letras del alfabeto inglés pronunciadas de forma aislada. Cada letra fue pronunciada y grabada 2 veces por cada uno de los 150 hablantes que intervinieron. Esto hace un total de, aproximadamente, 85 minutos de habla.

Los hablantes que participaron en las grabaciones tenían el inglés como lengua materna y una edad comprendida entre los 14 y los 72 años. La mitad eran hombres y la otra mitad mujeres.

4.1.2.2 DESCRIPCIÓN DEL ENTORNO DE EXPERIMENTACIÓN.

El entorno de experimentación que nosotros hemos creado para llevar a cabo nuestros experimentos se basa en el sistema de pruebas "ISOLET Testbed"[25], que implementa una aproximación al modelo híbrido de reconocimiento de voz.

Las distintas herramientas necesarias para montar este sistema de pruebas pueden obtenerse de "SPRACHcore-nogui", de "quicknet3", de ISOLET testbed y del conjunto de librerías " dpwelib-2009-02-24". Todas ellas pueden obtenerse de manera gratuita a través de la página web del International Computer Science Institute (ICSI)[26].

Para la preparación del entorno de experimentación, la base de datos ISOLET limpia fue contaminada con diferentes tipos de ruido para poder realizar los experimentos en condiciones menos ideales que las del laboratorio. Los ruidos fueron obtenidos de la colección ``RSG-10". Para llevar a cabo esta contaminación se siguieron tres principios fundamentales:

- Distintos tipos de ruido.
- Distintas Relaciones Señal a Ruido (SNR).
- Condiciones de entrenamiento limpia y ruidosa.

Para crear una base de datos con ruido [27] la base ISOLET se dividió en 5 partes del mismo tamaño para facilitar un procedimiento de validación cruzada, contaminándose todas ellas con tres

ruidos diferentes (``Speech babble", ``Factory floor noise 2", ``Car interior noise'').  Y a continuación cada una de las 5 partes con 1 de los siguientes ruidos (``Pink noise", ``F-16 cockpit noise", ``Destroyer operations room noise", ``Military vehicle noise", ``Factory floor noise 1''). Por cada una de estas partes existe un directorio con los diferentes archivos que codifican la voz.

Como ya hemos comentado en la sección anterior, la base de datos del habla está dividida en 5 subconjuntos limpios y otros 5 ruidosos. El orden en que los ficheros de las distintas secciones se lee es siempre el mismo y está especificado en los archivos ``noisy.wav.files.rand" y "clean.wav.files.rand", generado aleatoriamente..

Para llevar a cabo los experimentos se diseñaron distintos scripts que hicieron posible la parametrización de la base de datos (coeficientes MFCC y PLP), su posterior normalización y paso por el reconocedor.

Debemos tener en cuenta que esta herramienta está diseñada para usarse en sistemas operativos Unix; es por ello que para crear el entorno de experimentación fue necesario en primer lugar instalar en el hardware usado versiones @debian, en nuestro caso se escogió la distribución gratuita Ubuntu[28]. Tras esto se instalaron las herramientas necesarias para llevar a cabo los experimentos.

El hardware usado fue el siguiente:

Ordenador 1:

Procesador AMD Phenom™ II X4 995 Proccesor 803GHz

4GB RAM

Ordenador 2:

Intel® core™ 2 CPU T5500@1,66GHz

982MHz, 896MB RAM

Se usaron los dos ordenadores de forma simultánea para agilizar la realización de cada uno de los experimentos realizados, puesto que la duración de cada uno de ellos asciende a más de 15 horas.

4.1.3 SOBRE LA PREPARACIÓN DE LAS MATRICES DE CONFUSIÓN

Tal y como hemos relatado en el apartado 4.1.2 basaremos los análisis de los resultados de los experimentos RAH en los conjuntos de datos:

- ISOLET COMPLETO: Conjunto ISOLET sin reducir de dimensión 27x27
- ISOLET-M&N55: Subconjunto de alófonos comunes entre ISOLET y los experimentos de M&N de dimensión 11x11
- ISOLET VOCALES: Subconjunto de las vocales de dimensión 10x10
- ISOLET CONSONANTES: Subconjunto de las consonantes de la base de datos ISOLET de dimensión 17x17

Y los compararemos con los conjuntos de datos de los experimentos RHH siguientes:

- M&N55 COMPLETO: Conjunto de M&N 16x16
- M&N55-ISOLET: Subconjunto de alófonos comunes entre ISOLET y los experimentos de M&N 11x11

Debemos tener en cuenta que para la realización de las comparativas y análisis de los resultados obtenidos debido a la complejidad de nuestro entorno de experimentación y la imposibilidad de efectuar la reducción de datos antes de la realización de los experimentos, éstas reducciones se han llevado a cabo una vez finalizados los experimentos, es decir, se han extraído directamente las sub-matrices correspondientes de la matriz total (quitando la fila y columna correspondientes a la clase que se elimina). Esto hace que cambien ligeramente las probabilidades a priori de cada una de las clases.

Para cuantificar la importancia de esta modificación a continuación vamos a realizar un estudio de su influencia para que cuando posteriormente analicemos los resultados sepamos exactamente cuál es la repercusión en los Triángulos entrópicos y retículos de confusión representados.

Analizando las matrices de confusión resultantes de las reducciones podemos darnos fácilmente que el número de muestras de entrada en las matrices no será el mismo.

Para ello vamos a representar en primer lugar, de forma individual y después de forma conjunta los histogramas de las muestras de entrada de cada uno de los fonemas en las matrices de confusión resultantes en los siguientes casos de filtrado:

1. Situaciones más desfavorables: 200-300Hz y 4500-5000Hz
2. Situación más favorable: 200-6500Hz.

Separaremos el análisis representando por un lado los datos de entrada (fonemas) correspondientes al experimento RAH listados al principio de este apartado y por otro al experimento RHH (fonemas) MFCC tanto clean como noisy listados anteriormente.

A continuación pasamos a la representación de los histogramas RAH de cada uno de las tres situaciones descritas anteriormente comparando los 4 subconjuntos o submatrices de confusión:

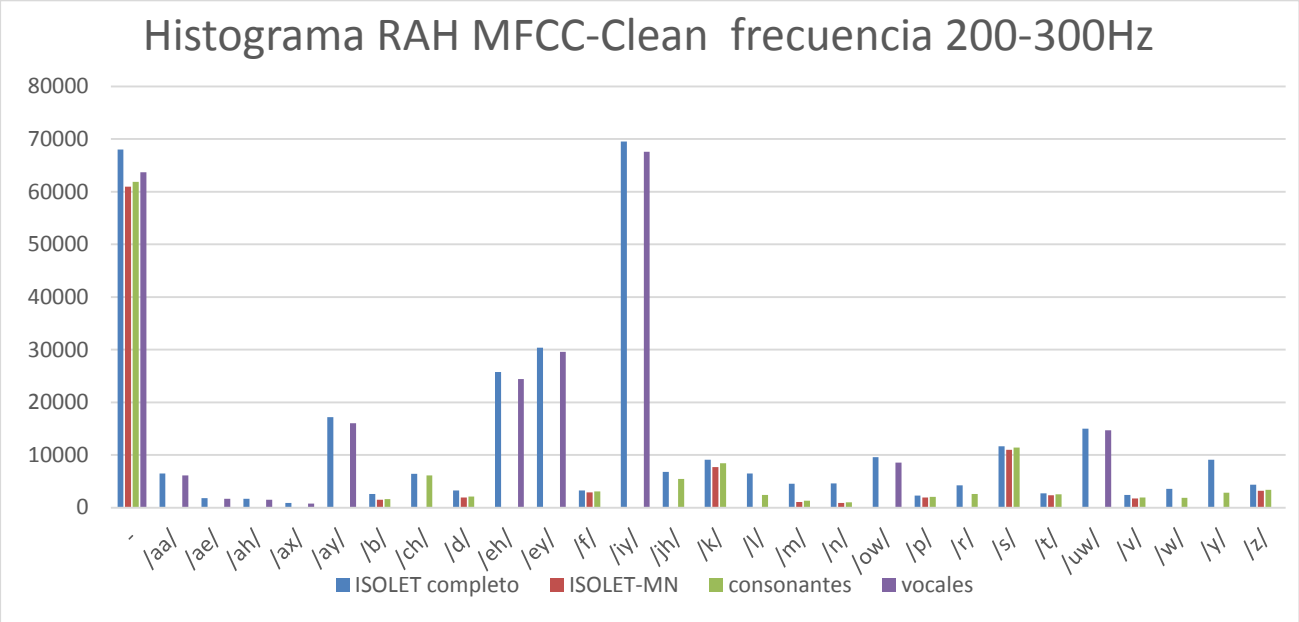

Figura 19; Histograma RAH MFCC-Clean para frecuencia 200-300Hz

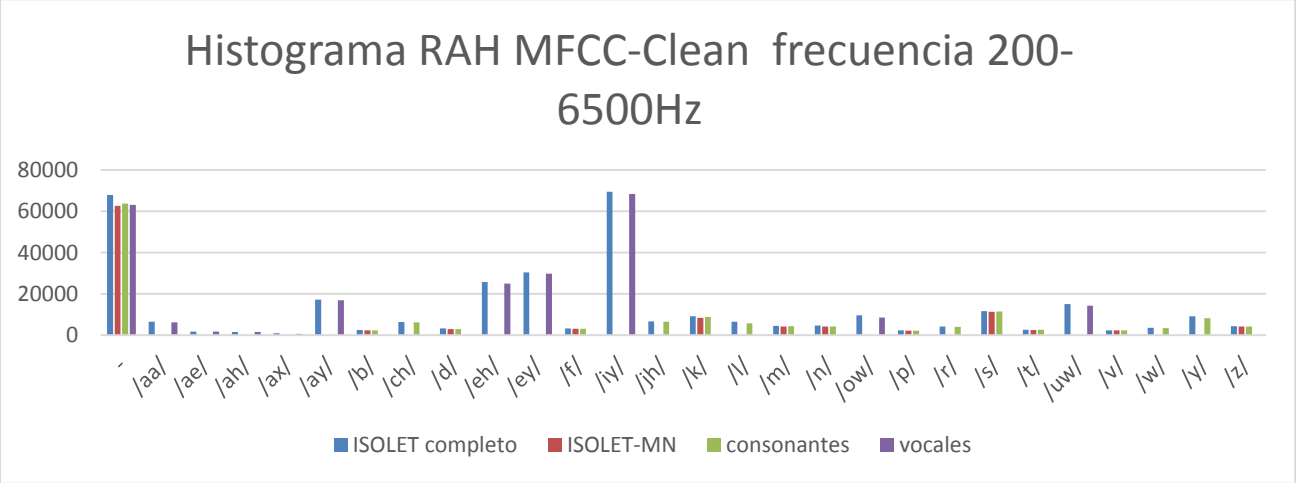

Figura 20: Histograma MFFC-Clean para frecuencia 200-6500Hz

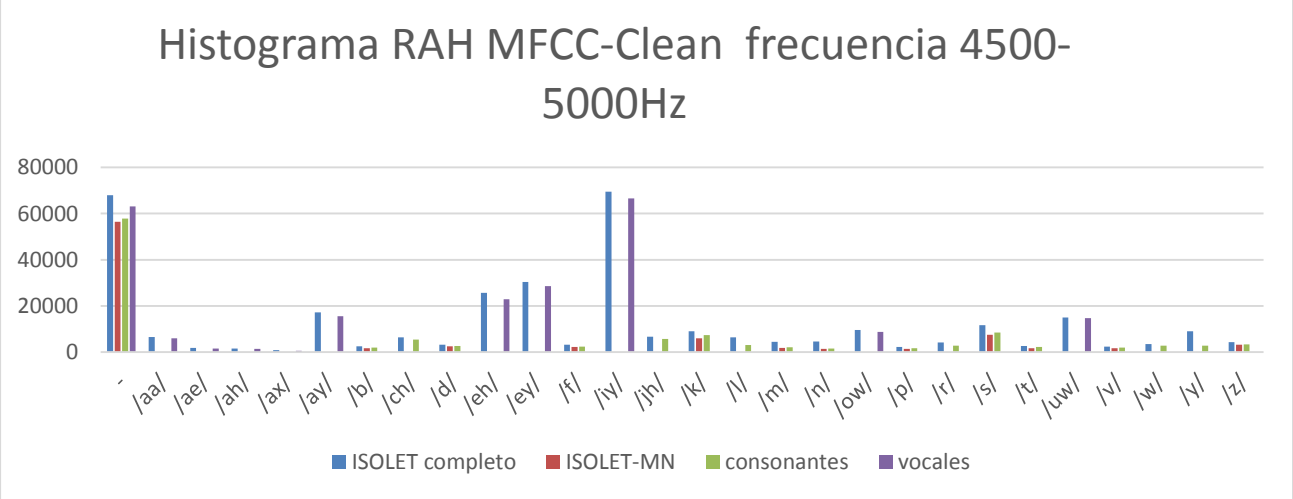

Figura 21: Histograma MFCC Clean para frecuencia 4500-5000Hz

Podemos observar las variaciones en el número de muestras de entrada (estímulos), que varía en función de la submatriz seleccionada siendo las variaciones más acusadas en las dos situaciones más desfavorables (200-300Hz y 4500-5000Hz) y casi inapreciables con el ancho de banda completo. Consideramos, en todo caso, que aún en los casos extremos, no debería representar un problema importante. Es necesario, sin embargo, introducir una salvedad: la de los fonemas nasales /m/ y /n/ en los que vemos que la diferencia entre el número total de muestras y las seleccionadas con los subconjuntos ISOLET-M&N55 e ISOLET CONSONANTES es muy grande y por tanto, debemos concluir que hay un gran número de confusiones entre estos dos fonemas y los vocálicos. Nuestra hipótesis para la explicación de este fenómeno es que la segmentación de estos fonemas es más difícil que en el resto y que se producen problemas de alineamiento.

A continuación, presentamos de forma conjunta las tres situaciones analizadas para cada uno de los 3 subconjuntos (ISOLET-M&N55, ISOLET CONSONANTES e ISOLET VOCALES).

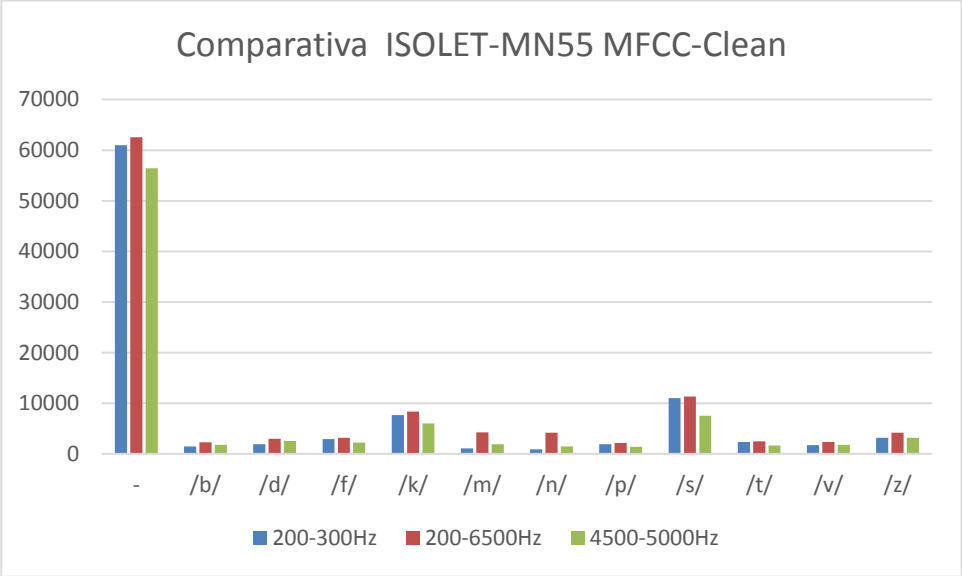

Figura 22: Histograma comparativo de ISOLET-M&N55 MFCC-Clean

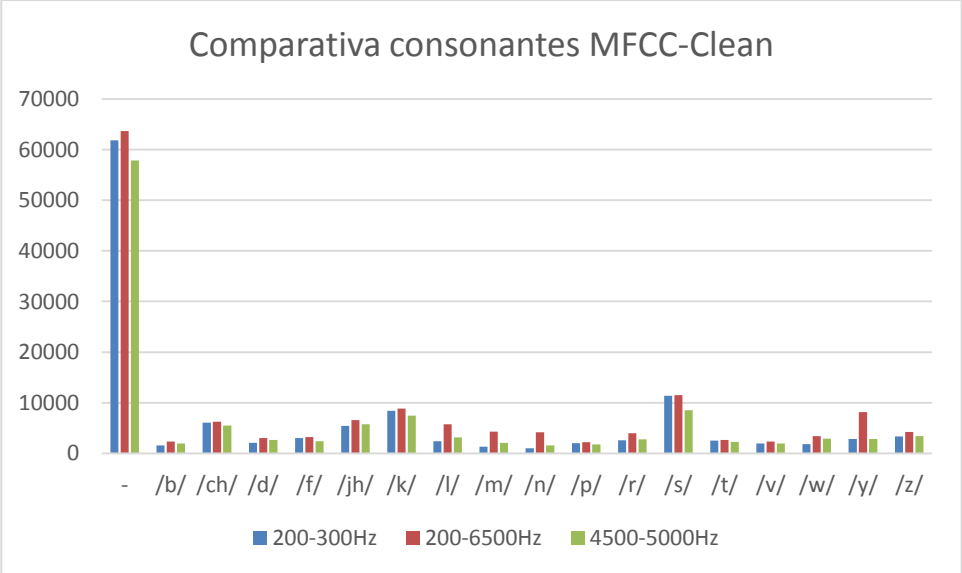

Figura 23: Histograma comparativo ISOLET CONSONANTES MFCC-Clean

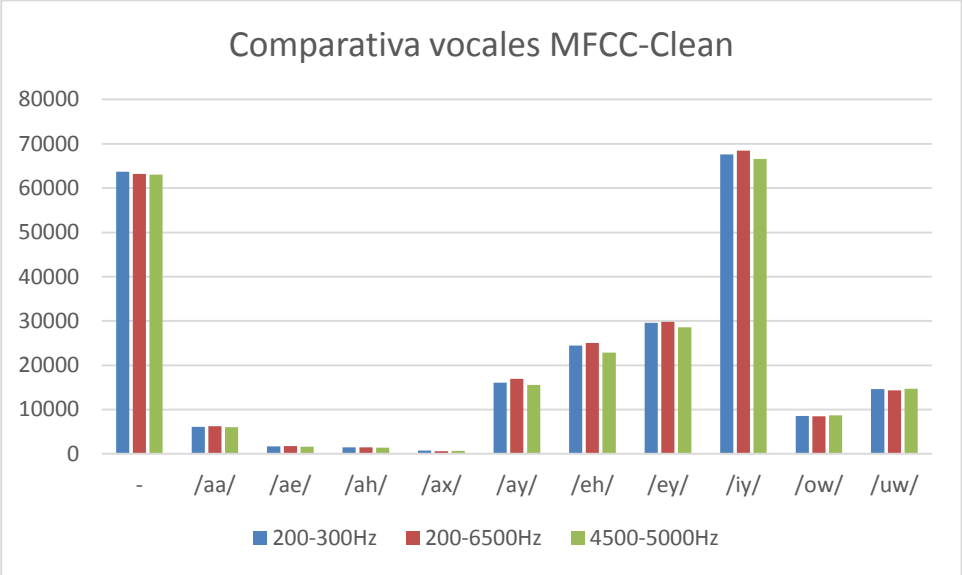

Figura 24: Comparativa histograma ISOLET VOCALES MFCC-Clean

Observando las figuras representadas nos damos cuenta de que la reducción de los datos de entrada en el análisis de nuestros experimentos no es elevada mostrándose en todo caso, de forma más acusada en las consonantes que en las vocales. Pensamos, por tanto, que no repercutirá de forma grave en nuestro análisis de los resultados aunque, eso sí, debemos tener en cuenta que nuestros datos de entrada en el triángulo entrópico se verán desplazados ligeramente hacía la derecha.

De nuevo con la excepción de las consonantes nasales, vemos que el número de muestras de entrada no se ve reducido de forma notable por el proceso de selección de las submatrices de confusión, haciendo por tanto validos nuestros análisis posteriores.

Repetimos ahora el mismo análisis con la base de datos ruidosa.

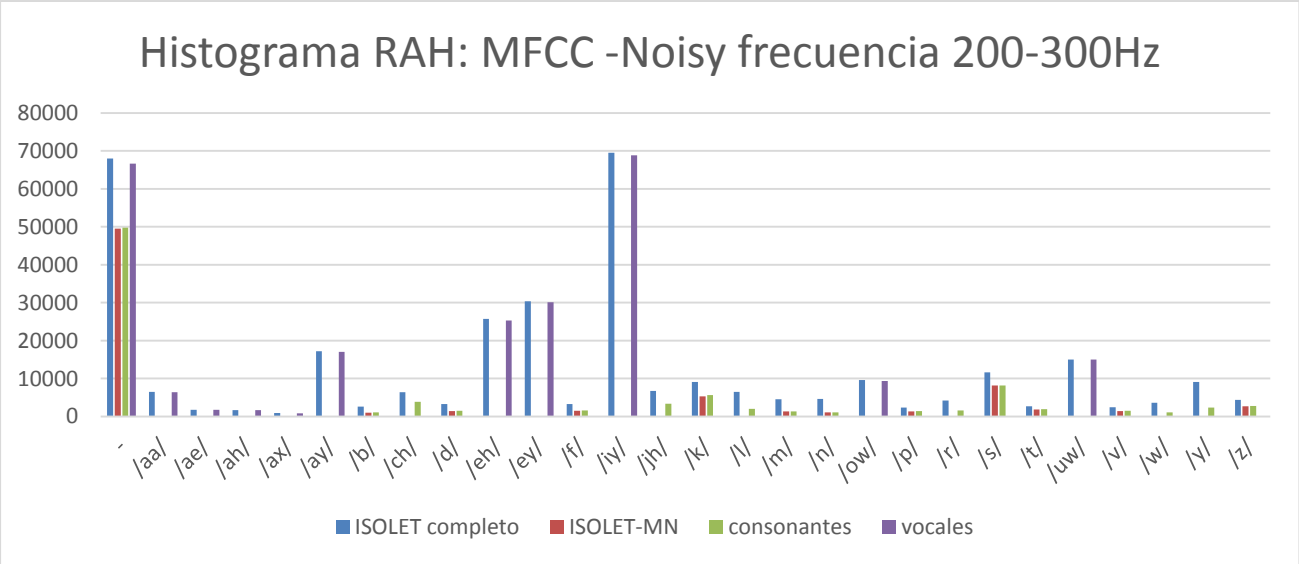

Figura 25: Histograma MFCC Noisy para frecuencia 200-300 Hz

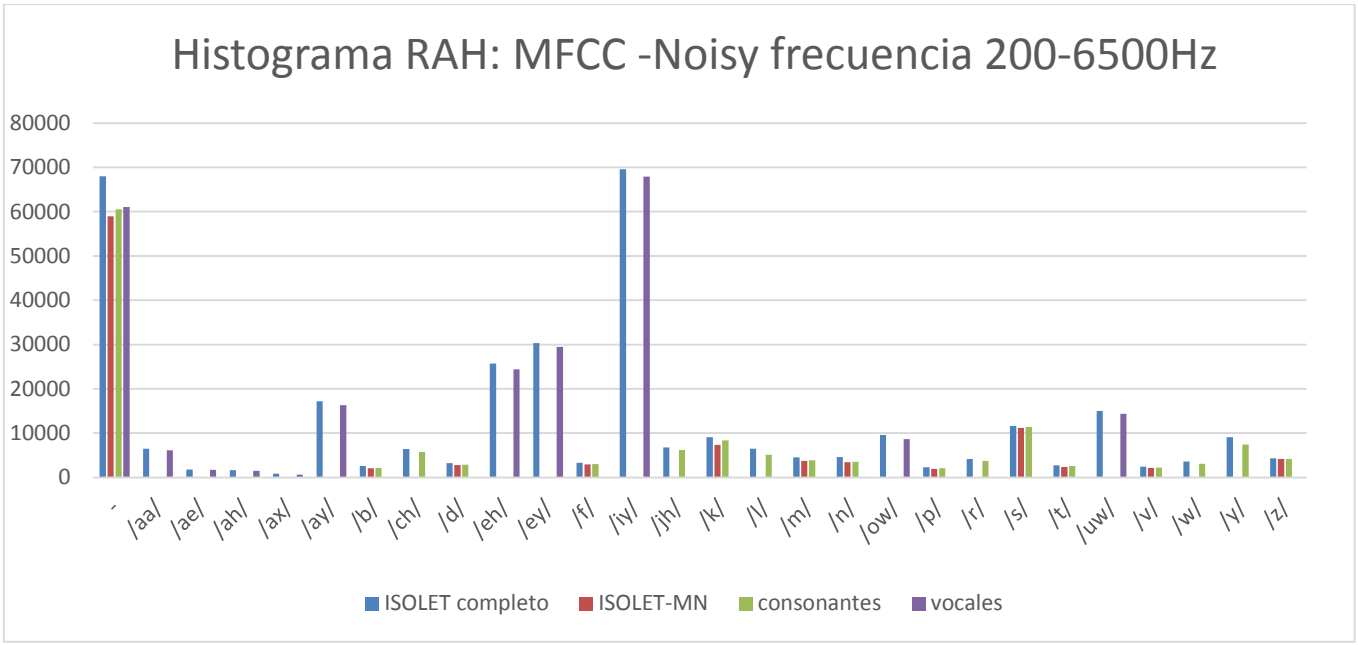

Figura 26: Histograma MFCC Noisy para frecuencia 200-6500Hz

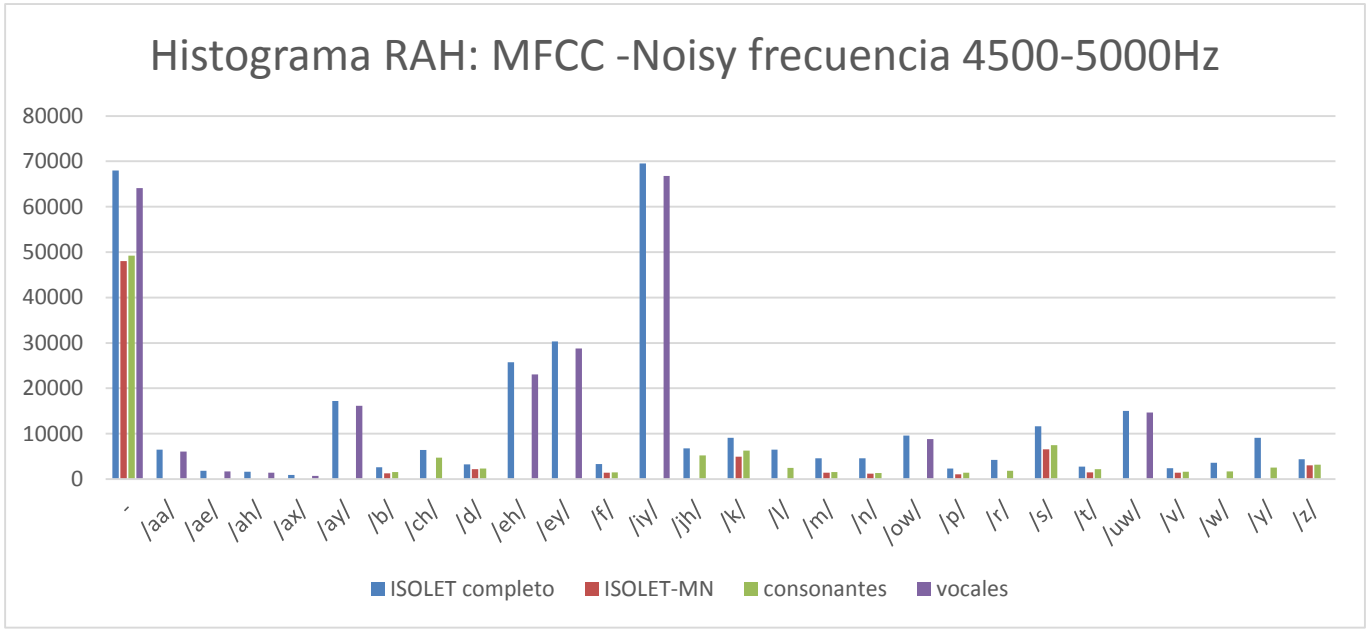

Figura 27: Histograma MFCC Noisy para frecuencia 4500-5000Hz

En este caso observamos las mismas tendencias que ya vimos en el caso no ruidoso más agravadas pero nos llama la atención el comportamiento del silencio (representado por " –") que vemos reducirse notablemente en las dos situaciones más adversas (200-300Hz y 4500-5000Hz) para los subconjuntos ISOLET-M&N55 e ISOLET CONSONANTES, probablemente debido a los errores de segmentación causados por una dificultad mayor en delimitar el silencio y las consonantes en un ambiente ruidoso que sin embargo, es menos acusado en el caso de las vocales.

De nuevo, presentamos de forma conjunta las tres situaciones analizadas para cada uno de los 3 subconjuntos (ISOLET-M&N55, ISOLET CONSONANTES e ISOLET VOCALES) esta vez en los casos ruidosos.

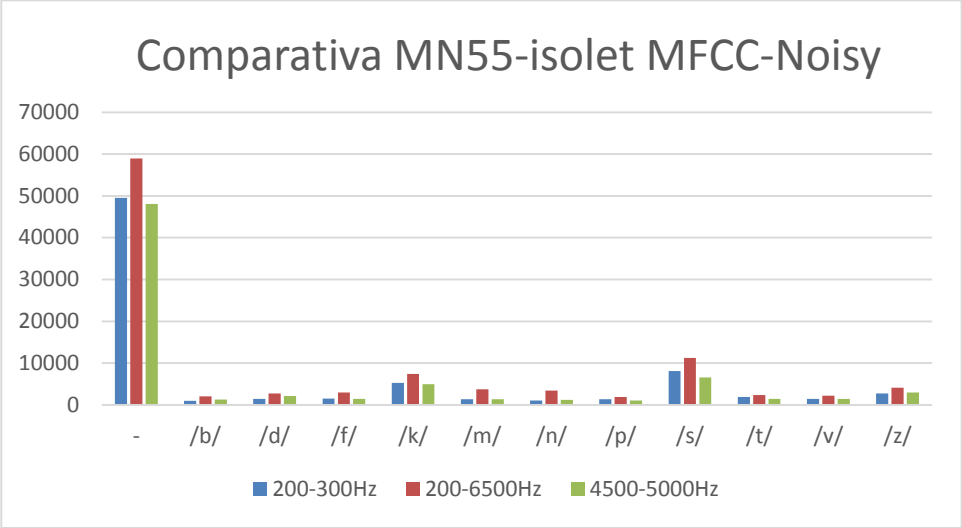

Figura 28: Comparativa histograma ISOLET-M&N55 MFCC-Noisy

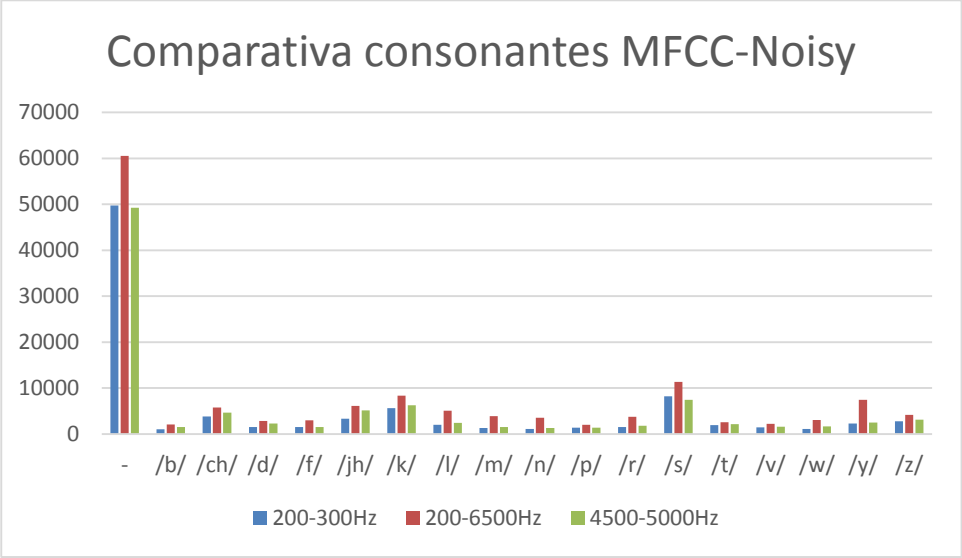

Figura 29: Comparativa histograma ISOLET CONSONANTES MFCC-Noisy

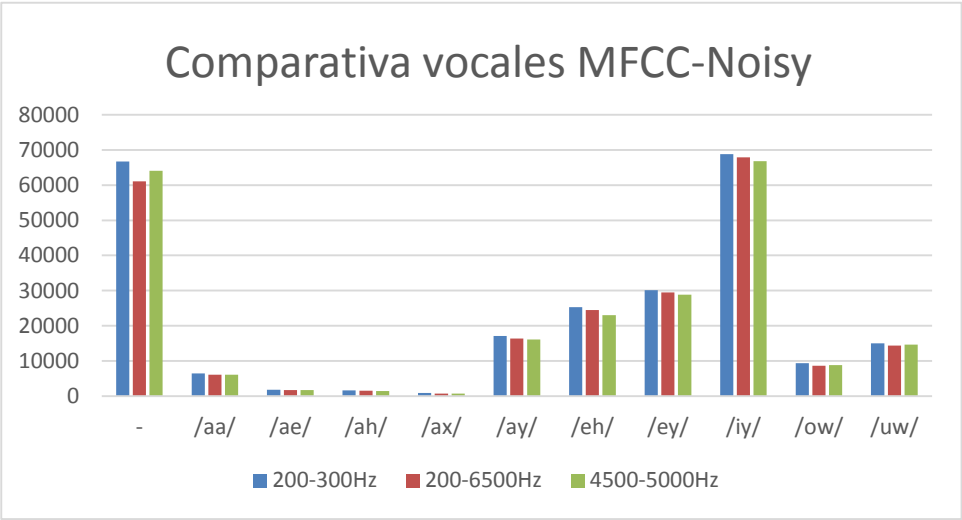

Figura 30: Comparativa histograma ISOLET VOCALES MFCC-Noisy

Podemos concluir que las leves modificaciones que observamos en los datos sin ruido son ahora más notables y aunque consideramos que nuestro análisis posterior seguirá siendo válido debemos tener en cuenta este hecho por si pudiera tener alguna influencia.

4.2 ANÁLISIS DE LOS RESULTADOS.

Una vez que hemos realizado los experimentos pasamos a la parte más importante de nuestro estudio que es analizar los resultados obtenidos. Para ello utilizaremos las herramientas introducidas en el capítulo 3 de este documento.

Para poder analizar los resultados obtenidos en la ejecución de nuestros experimentos debemos preparar los datos obtenidos de tal forma que puedan ser analizados mediante las herramientas anteriormente descritas.

En particular, hemos obtenido y preparado las matrices de confusión, tanto en RHH como en RAH, usando éstas matrices como las variables de entradas para las distintas funciones programadas en matlab para la obtención y representación de nuestras herramientas de análisis: los triángulos entrópicos y las matrices de incidencia que después convertiremos en retículos de confusión utilizando la herramienta CONEXP[29].

Dada la gran cantidad de matrices de confusión que hemos obtenido realizaremos primero el análisis con los triángulos entrópicos y basándonos en ellos seleccionaremos aquellos casos más significativos para su posterior análisis con los retículos de confusión.

4.2.1 ANÁLISIS DEL EXPERIMENTO RHH

En este apartado analizaremos los resultados del experimento RHH, los datos que vamos a representar corresponden a los resultados de los experimentos de M&N55.

Las frecuencias de corte analizadas y los subconjuntos de datos representados han sido descritos con anterioridad en los apartados 2 y 4 de este documento.

Analizaremos primeramente los Triángulos entrópicos y una vez analizados éstos escogeremos las frecuencias más interesantes para el análisis de los retículos de confusión.

4.2.1.1 Triángulos entrópicos

En este apartado representaremos los Triángulos entrópicos basándonos en dos factores, primeramente representación en el que el color representa la precisión y seguidamente otra en la que dicho color representa la frecuencia de corte.

Para la representación separaremos los que denominamos el análisis en frecuencias de paso alto y paso bajo.

Debemos de tener en cuenta para la representación basada en la frecuencia que en los triángulos para una visualización más clara se representan las frecuencias en escala logarítmica log (frec (hz)/1000) quedando las frecuencias representadas de la siguiente manera:

Paso bajo:

Frecuencia corte inferior	Frecuencias corte superior
200 frec en Hz	[300,400,600,1200,2500,5000,6500]
-0.6990 frec en log	[-0.5229,-0.3979,-0.2218, 0.0792, 0.3979, 0.6990, 0.8129]

Paso alto:

Frecuencias corte inferior	Frecuencia corte superior
[1000,2000,2500,3000,4500]	5000 frec en Hz
[0, 0.3010, 0.3979, 0.4771 0.6532]	0.6990 frec en log

Para más información sobre la representación y análisis de los Triángulos entrópicos visitar apartado 3.2 de este documento.

4.2.1.1.1 Representación paso bajo

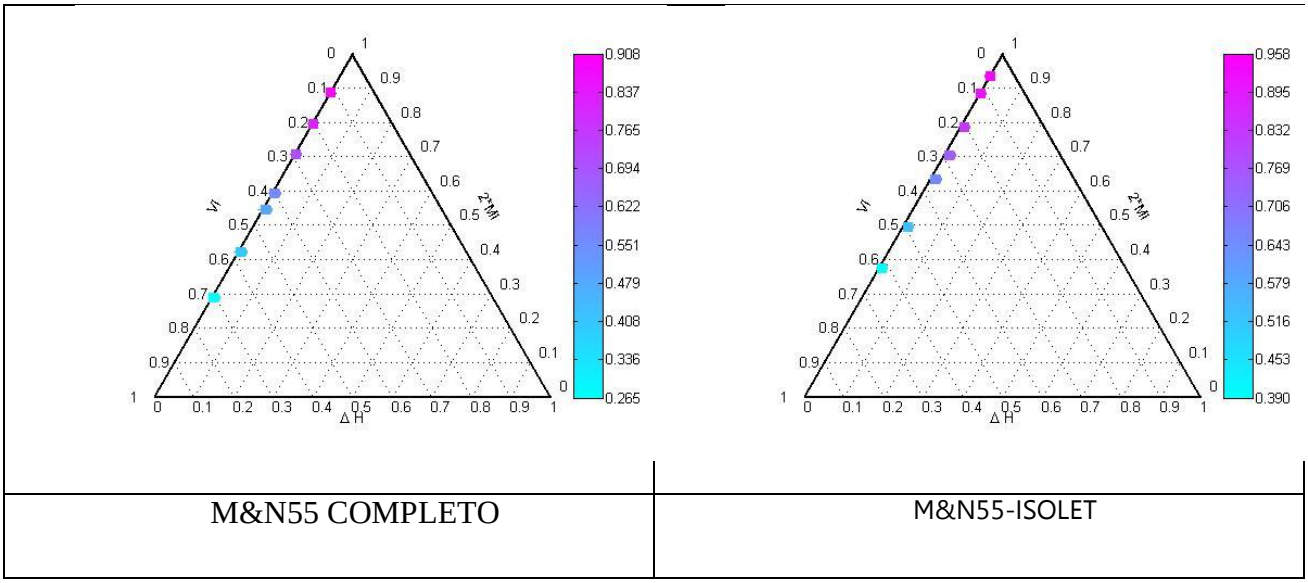

Figura 31: triángulos entrópicos RHH paso bajo con representación de la precisión.

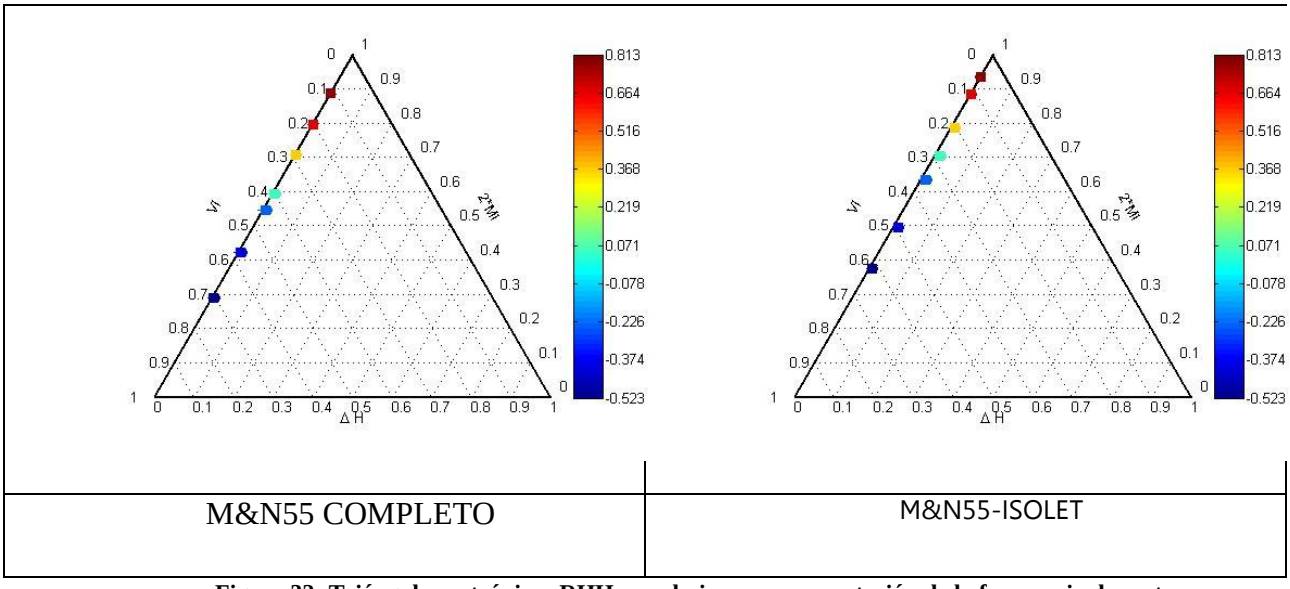

Figura 32: Triángulos entrópicos RHH paso bajo con representación de la frecuencia de corte

Observando los Triángulos entrópicos representados en la Figura 31 y la Figura 32vemos que:

En el caso de RHH para las frecuencias de paso bajo los resultados tanto para los datos totales del experimento M&N55 como para los alófonos comunes entre ISOLET y M&N55 se encuentran en la zona más a la izquierda del triángulo llevando a cabo una progresión hacia el punto más alto del mismo donde todos los datos (tanto entrada como salida) están alineados, como cabía esperar puesto que las locuciones de entrada están equilibradas (existe el mismo número de ejemplos para todas las clases). Esto indica que M&N55 colocaron a los sujetos del test en la condición más difícil posible impidiendo que acertaran haciendo uso de información sobre la frecuencia de aparición de los fonemas.

Observamos además que a frecuencias de corte más restrictivas el reconocedor obtiene peores resultados, más cercanos a la esquina inferior izquierda, y según mejoran las frecuencias de corte progresa hacia la zona superior izquierda del triángulo lo que demuestra que en los peores casos el reconocedor obtiene los mejores resultados.

Si comparamos los triángulos obtenidos para todos los fonemas del experimento de M&N55 y los de la intersección con los de ISOLET llegamos a la conclusión de que este último subconjunto es algo más sencillo pues los resultados para cada una de las frecuencias de corte se sitúan en una zona superior del triángulo. Es necesario recordar que la selección de los fonemas de este subconjunto se hace "a posteriori" y por lo tanto no influye en los resultados individuales que siguen siendo los mismos, pero debido a que cuando eliminamos una fila de la matriz de confusión eliminamos también su correspondiente columna, todos los errores de los fonemas emitidos que no se eliminan hacia fonemas recibidos fuera del subconjunto seleccionado ya no se tienen en cuenta.

4.2.1.1.2 Representación paso alto

A continuación los correspondientes paso alto:

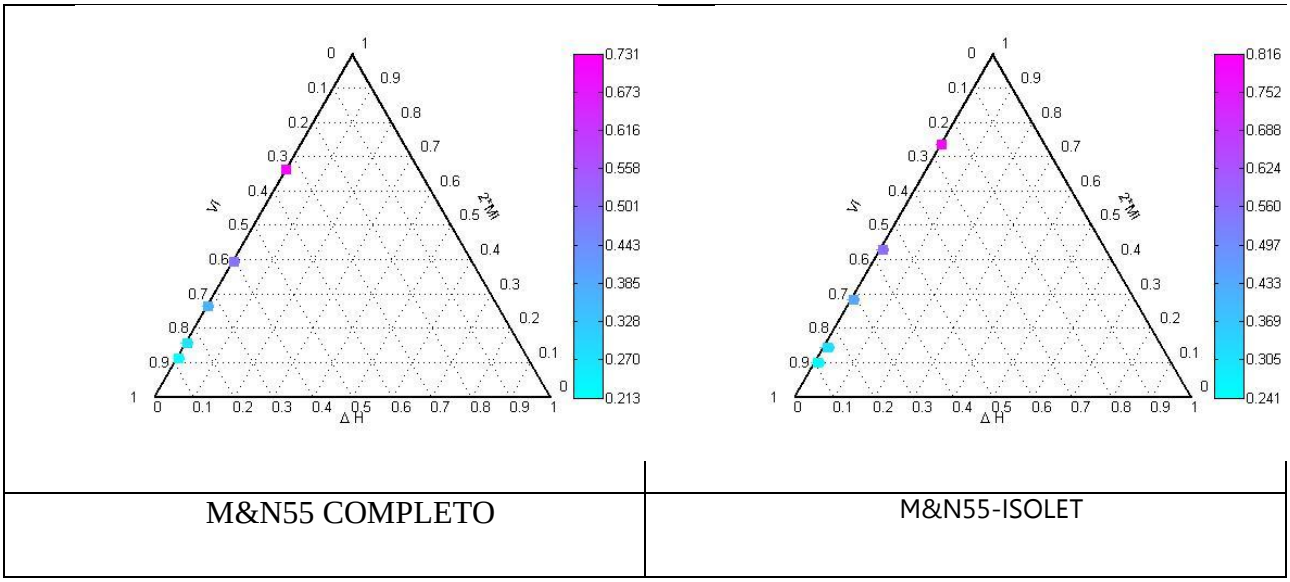

Figura 33: Triángulos entrópicos RHH paso alto con representación de la precisión

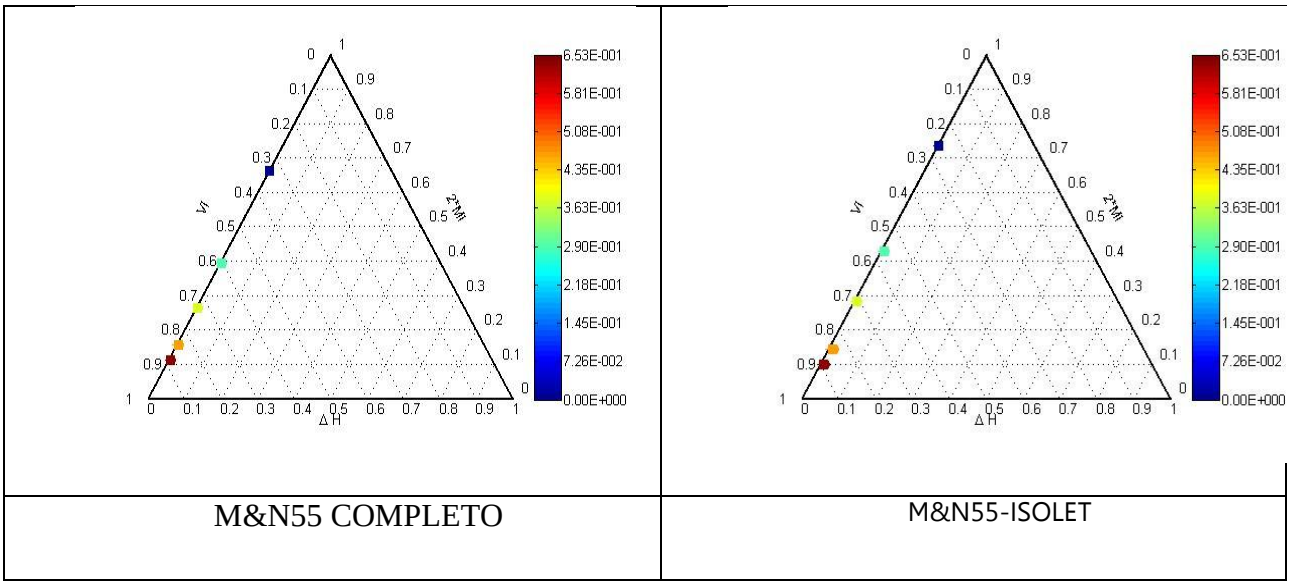

Figura 34: Triángulos entrópicos RHH paso alto con representación de la frecuencia de corte

Analizando los Triángulos entrópicos representados en la Figura 33 y la Figura 34:

En el caso de RHH para las frecuencias de paso alto los resultados tanto para los datos totales del experimento M&N55 como para los alófonos comunes entre ISOLET y M&N55 se encuentran en la zona más izquierda del triángulo llevando a cabo una progresión hacia el punto más bajo del mismo donde todos los datos (tanto entrada como salida) están alineados, esto nos dice que el reconocedor está obteniendo buenos resultados.

Observamos además que a frecuencias de corte más restrictivas el reconocedor obtiene peores resultados, más cercanos a la esquina inferior izquierda siendo mejores los resultados donde el rango de frecuencia utilizado es menos restrictivo.

Teniendo en cuenta los resultados del análisis del caso de paso bajo mencionados anteriormente nos damos cuenta que este reconocedor funciona mejor a frecuencias bajas ya que los resultados obtenidos por el mismo en el peor de los casos de paso bajo son mejores que los obtenidos en el peor de los casos del estudio paso alto.

4.2.1.2 *Retículos de confusión*

Una vez estudiados los triángulos entrópicos pasamos a analizar de forma más pormenorizada las frecuencias con mayor interés representadas anteriormente.

Para este estudio obtendremos los retículos de confusión e interpretaremos los mismos.

Para el caso del experimento RHH vamos a representar los retículos correspondientes a las siguientes matrices de confusión:

Matrices de confusión filtradas a: 200-600 Hz, 200-1200Hz, 200-5000Hz y 2000-5000Hz

Debemos tener en cuenta que para la representación de los retículos solo usamos los fonemas de cada subconjunto de datos que se confunden entre sí, aquellos en los que no hay confusión no los representamos en el retículo de confusión. De esta forma obtenemos retículos de confusión simplificados en los cuales, las etiquetas ausentes implican que no hay confusiones de dicho fonema (para el nivel de análisis escogido φ).

No hemos de olvidar que para cada retículo representamos el número de conceptos que hace posible un análisis claro. Esto lo hacemos posible seleccionando un valor de φ adecuado que determinará el número de conceptos, éste será independiente en cada retículo. Un número muy alto de conceptos representa demasiadas confusiones lo que hace de la representación y el análisis una tarea bastante compleja y que no aporta demasiado y uno excesivamente bajo no nos muestra confusiones por tanto el análisis carece de interés. Por tanto elegimos para cada retículo el número de conceptos que hacen posible analizarlo correctamente.

4.2.1.2.1 Frecuencia 200-600Hz

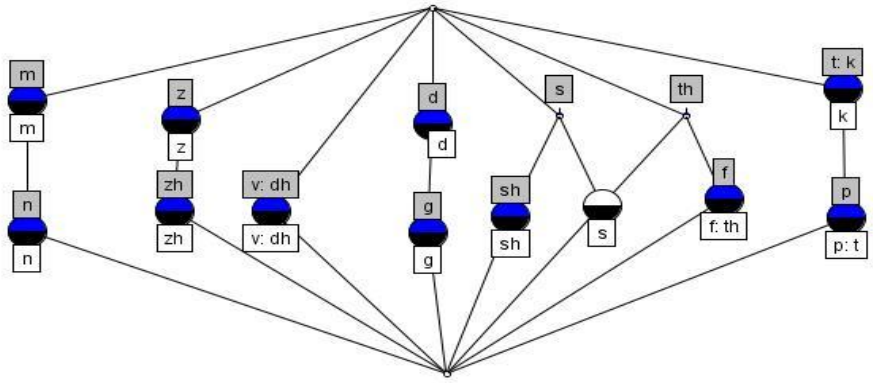

Figura 35: Retículo de confusión del experimento M&N55 COMPLETO con φ=-1.948367 y 16 conceptos.

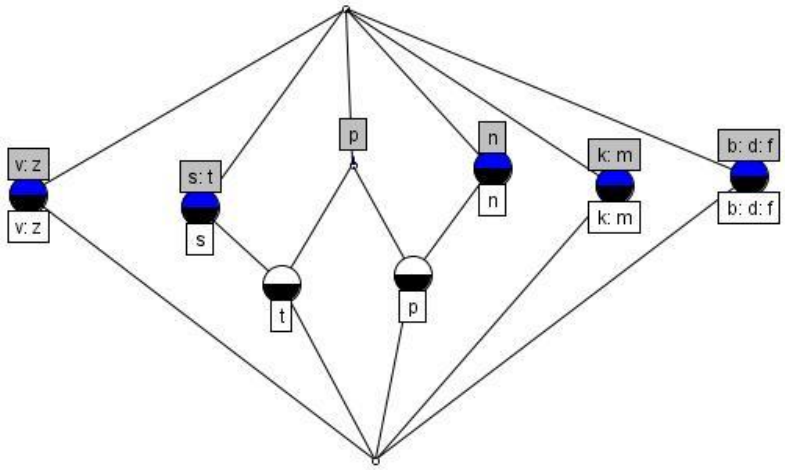

Figura 36: Retículo de confusión del subconjunto M&N55-ISOLET con φ= 0.170133 y 10 conceptos.

En el caso del retículo de confusión representado para el experimento M&N55 COMPLETO (figura 35) podemos ver claramente que para la frecuencia seleccionada las confusiones se agrupan de la siguiente manera:

1) /m/y /n/; /p/, /t/, /k/; /zh/ y/ z/; /v/ y /dh/; / d/ y /g/ definen 5 grupos de fonemas que sólo se confunden entre ellos mismos estando separados del resto, es decir, forman sub-retículos adjuntos cuyos únicos elementos en común son el "top"· (el supremo) y el "bottom" (el ínfimo). De estos conjuntos /m/ y /n/ se engloban en los denominados en [16] como conjuntos de fonemas más fáciles de confundir por sus similitudes fonéticas y corresponden con los fonemas nasales, /p/, /t/ y /k/ son los fonemas oclusivos sordos, /zh/ y /z/ comparten el ser fricativas sonoras (postalveolar y alveolar, respectivamente), /v/ y /dh/ son también fricativas sonoras (dental y labioldental, respectivamente y

/d/ y /g/ son oclusivas sonoras aunque notamos en este grupo la ausencia de /b/ (ausente en el retículo debido a que no se confunde con ninguna más).

2) Sin embargo observamos cómo /sh/ solo se confunde con /s/, mientras que /s/ a su vez se confunde con / f/ y /th/ para el experimento 200-600Hz que es el ejemplo de frecuencia de corte paso bajo reducido que hemos representado. Es llamativo cómo a pesar de lo restrictivo de estas frecuencias de paso se mantiene una estructura de confusiones coherente con las características articulatorias clásicas mencionadas.

Si pasamos a observar para esta misma frecuencia el retículo de confusión representado para el conjunto de los alófonos comunes entre M&N55 e ISOLET (figura 36) observamos que las confusiones se agrupan de la siguiente manera:

1) /k/y /m/; /v/ y /z/; /b/, /d/ y /f/ definen 3 grupos de fonemas que sólo se confunden entre ellos mismos estando separados del resto, es decir, forman sub-retículos adjuntos. De estos conjuntos /v/ y /z/ se engloban en los denominados en [21] como conjuntos de fonemas más fáciles de confundir por sus similitudes fonéticas y comparten ser fricativas labio-dentales y alveolares respectivamente. /b/, /d/ y /f/ donde /b/ y /d/ son fonemas oclusivos sordos mientras que /f/ es fricativo labio-dental y en el conjunto formado por /k/ y /m/ que no tienen nada en común ya que /m/ es nasal y /k/ oclusiva velar.

2) Sin embargo observamos cómo el conjunto /p/ y /t/ que son fonemas oclusivos sordos pertenecientes a [21] no se confunden únicamente entre ellos sino que /p/ se confunde a su vez con el fonema nasal /n/ y que este a su vez se confunde con el fonema fricativo alveolar /s/.

El hecho de haber restringido el número de fonemas observados nos ha permitido entrar en más profundidad a observar las confusiones y debido a ello y a diferencia con el caso anterior analizado vemos como para estas frecuencias tan restrictivas aparecen ya algunas incoherencias con las características articulatorias clásicas mencionadas. Específicamente, las consonantes nasales no aparecen como un sub-retículo independiente sino relacionada con otros fonemas, en un principio con características acústico-articulatorias muy diferentes.

4.2.1.2.2 Frecuencia 200-1200Hz

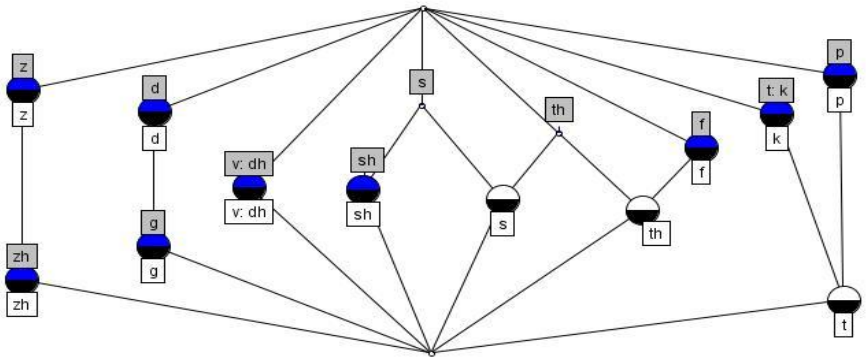

Figura 37: Retículo de confusión del experimento M&N55 COMPLETO con φ=-1.99216 y 16 conceptos.

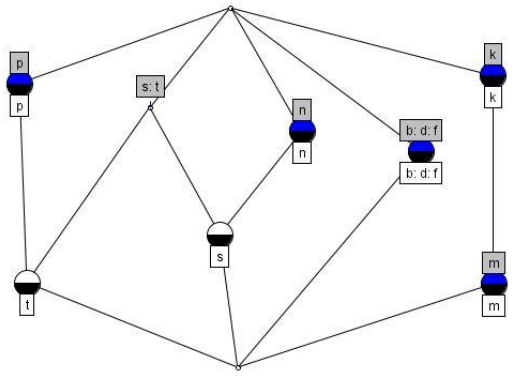

Figura 38: Retículo de confusión del subconjunto M&N55-ISOLET con φ= 0.229252 y 10 conceptos.

En las figuras 37 y 38 que son las correspondientes a la frecuencia 200-1200Hz podemos ver que al ampliar el rango de frecuencias las confusiones entre los fonemas se agrupan más fácilmente, y que cuanto mayor es el ancho de banda de paso mejor se distingue la señal. En el caso del retículo de confusión representado para el experimento M&N55 COMPLETO (figura 37) podemos ver claramente que para la frecuencia seleccionada las confusiones se agrupan de la siguiente manera:

1) /d/ y /g/; /zh/ y/ z/; /v/ y /dh/; /p/, /t/ y /k/ definen 4 grupos de fonemas que sólo se confunden entre ellos mismos estando separados del resto, es decir, forman sub-retículos adjuntos. El conjunto formado por /zh/ y /z/ comparten el ser fricativas sonoras (pos alveolar y alveolar, respectivamente), /v/ y /dh/ son también fricativas sonoras (dental y labio-dental, respectivamente y /d/ y /g/ son oclusivas sonoras aunque notamos en este grupo la ausencia de /b/ (ausente en el retículo debido a que no se confunde con ninguna más). También notamos que esta vez /p/ y /t/ se confunden entre sí y además /t/ se confunde con /k/ teniendo estos tres fonemas en común ser oclusivos sordos.

2) Un quinto sub-retículo adjunto más complejo aparece formado por /sh/, /s/ y /th/. Observamos cómo /sh/ solo se confunde con /s/, mientras que /s/ a su vez se confunde con /th/ y esta misma con /f/ que tienen en común ser fonemas fricativos.

Si pasamos a observar para esta misma frecuencia el retículo de confusión representado para el conjunto de los alófonos comunes entre M&N e ISOLET (figura 38) observamos que las confusiones se agrupan de la siguiente manera:

1) /k/y /m/; /b/, /d/ y /f/ definen 2 grupos de fonemas que sólo se confunden entre ellos mismos estando separados del resto, es decir, forman sub-retículos. De estos conjuntos /b/ y /d/ son los fonemas oclusivos sonoros mientras que /f/ es fricativo labio-dental y en el conjunto formado por /k/ y /m/ que no tienen nada en común ya que /m/ es nasal y /k/ oclusiva velar.

2) Sin embargo observamos cómo el conjunto /p/ y /t/ que son fonemas oclusivos sordos no se confunden únicamente entre ellos sino que /t/ se confunde a su vez con el fonema fricativo /s/ y que te a su vez se confunde con el fonema nasal /n/. En este caso vemos un comportamiento similar en la figura 36 analizada anteriormente y que ya hemos comentado.

4.2.1.2.3 Frecuencia 200-5000Hz

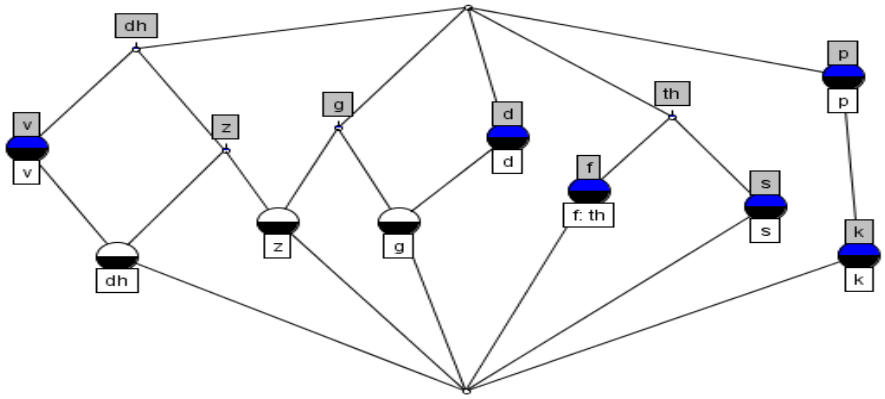

Figura 39: Retículo de confusión del experimento M&N55 COMPLETO con φ= -0.404507 y 15 conceptos.

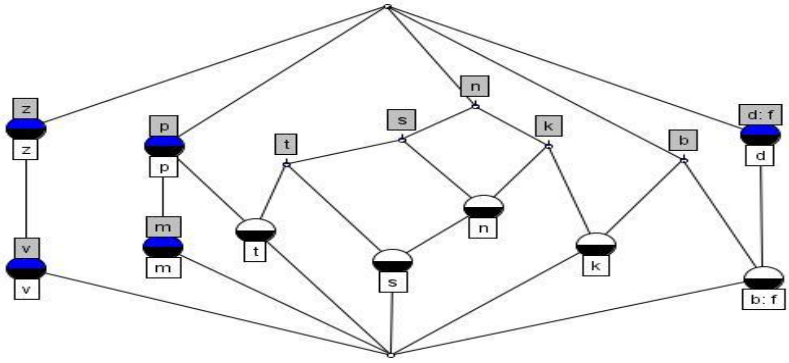

Figura 40: Retículo de confusión del subconjunto M&N55-ISOLET con φ= 3.157 y 17 conceptos.

En las figuras 39 y 40 nos encontramos en el mejor de los casos, podríamos considerar que a estas frecuencias el experimento no está filtrado, por tanto debería en el que mejores resultados se obtienen, en el análisis anterior de los triángulos entrópicos era así, pero ¿Cómo se confunde el clasificador?

En el caso del retículo de confusión representado para el experimento M&N55 COMPLETO (figura 39) podemos ver claramente que para la frecuencia seleccionada las confusiones se agrupan de la siguiente manera:

1) /p/ y /k/ define un conjunto de fonemas que sólo se confunden entre ellos mismos estando separados del resto, es decir, forman sub-retículos y que comparten el ser oclusivas sordas.

2) Podemos observar como en este caso el conjunto formado por /m/ y /n/ no aparece en la representación lo que significa que están perfectamente clasificados y no se confunden entre sí.

3) Del resto de fonemas que aparecen en el retículo podemos decir que se establecen bastantes confusiones entre ellos como pueden ser /s/ con /th/ y /f/ con /th/ siendo fonemas oclusivos todos ellos.

4) También observamos cómo /v/ fonema fricativo se confunde con /dh/ (fricativo) confundiéndose a su vez /dh/ con /z/ fonema fricativo.

5) Los fonemas oclusivos /g/ y /d/ se confunden entre sí, además /g/ se confunde con el fonema fricativo /z/.

Si pasamos a observar para esta misma frecuencia el retículo de confusión representado para el conjunto de los alófonos comunes entre M&N55 e ISOLET (figura 40) observamos que las confusiones se agrupan de la siguiente manera:

1) /v/ y /z/; define un conjunto de fonemas que sólo se confunden entre ellos mismos estando separados del resto, es decir, forman sub-retículos adjuntos. El conjunto formado por /v/ y /z/ comparten el ser fricativas y pertenecientes a [21].

2) Por otro lado observamos como el fonema fricativo /f/ se confunde con los oclusivos /d/ y /b/ y que este último se confunde a su vez con /k/ (oclusivo).

3) Las confusiones formadas por /m/ fonema nasal que se confunde con /p/ que es oclusivo es notoria, sí cabría esperar las confusiones de /t/ con /p/ y /s/ al tener los 3 características articulatorias similares.

Vemos que aunque la frecuencia de corte sea menos restrictiva siguen existiendo confusiones y con alguna salvedad se asemejan mucho a las características articulatorias mencionadas anteriormente.

4.2.1.2.4 Frecuencia 2000-5000Hz

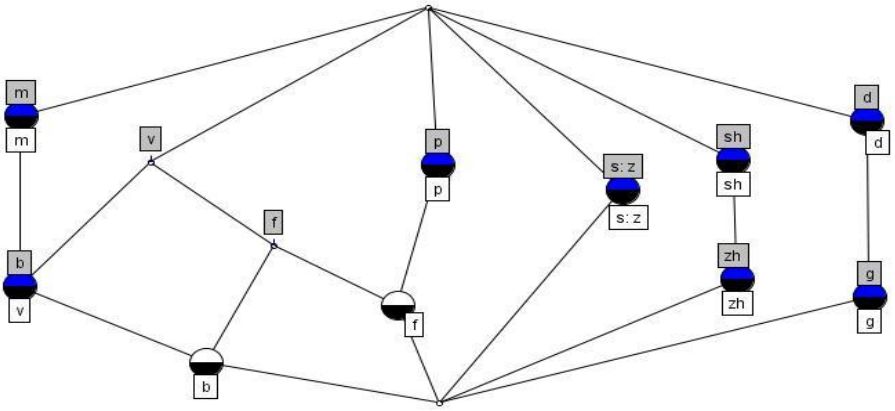

Figura 41: Retículo de confusión del experimento M&N55 con φ= -1.02127 y 14 conceptos.

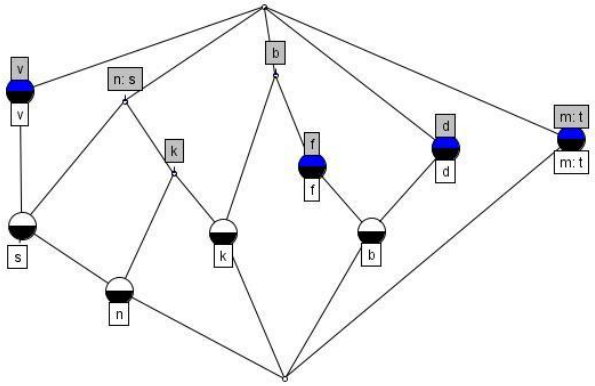

Figura 42: Retículo de confusión del subconjunto M&N55-ISOLET con φ= -0.230128 y 13 conceptos.

En las figuras 41 y 42 podemos observar:

Estas figuras corresponden al peor de los casos en filtrado paso alto, podemos observar cómo hemos ido viendo a lo largo del análisis del clasificador RHH que el comportamiento del mismo en torno a la confusión es más sencillo y con menos dependencias.

En el caso del retículo de confusión representado para el experimento M&N sin reducir (figura 41) podemos ver claramente que para la frecuencia seleccionada las confusiones se agrupan de la siguiente manera:

1) /d/ y /g/; /sh/ y /zh/; y /s/ y /z/ definen 3 grupos de fonemas que sólo se confunden entre ellos mismos estando separados del resto, es decir, forman sub-retículos adjuntos. De estos conjuntos el formado por /d/ y /g/ comparten el ser oclusivos, /s/ y/sh/ son fricativas y por último /s/ y /z/ son fricativas alveolares ambas.

2) El resto de los fonemas forman el cuarto sub-retículo adjunto con una complejidad mucho mayor en sus confusiones: vemos que /f/ (fricativa) se confunde con /p/ y /v/, además la /b/ se confunde con la /v/ siendo /b/ oclusiva y /v/ fricativa y a su vez con /m/ que es nasal y articulatoriamente bastante diferente. Podemos concluir que la falta de la banda de frecuencias inferior produce confusiones entre las fricativas y las oclusivas que antes no existían así como confusiones entre consonantes sonoras y sordas que no habíamos observado con el ancho de banda completo.

Si pasamos a observar para esta misma frecuencia el retículo de confusión representado para el conjunto de los alófonos comunes entre M&N55 e ISOLET (figura 42) observamos que las confusiones se agrupan de la siguiente manera:

1) /m/y /t/; define un conjunto de fonemas que sólo se confunden entre ellos mismos estando separados del resto, es decir, forman un sub-retículos adjunto al de las demás consonantes. El conjunto formado por /m/ y /t/ no tienen nada en común ya que /m/ es nasal y /t/ es oclusiva. Este tipo de confusiones puede deberse al encontrarnos en una frecuencia muy restrictiva.

2) Observamos cómo /n/ (nasal) se confunde con /k/ (oclusiva), /s/ y /v/ que son fricativas.

3) Volvemos a encontrar relaciones de confusión entre /b/, /d/ y /f/ que ya nos habíamos encontrado en el experimento sin reducción de ancho de banda.

4.2.2 ANÁLISIS DEL EXPERIMENTO RAH

Vamos a proceder de forma análoga al apartado 4.2.1 de este documento. Es decir, comenzaremos con los triángulos entrópicos y después proseguiremos con los retículos de confusión para algunas de los anchos de banda que seleccionaremos.

4.2.2.1 Triángulos entrópicos

En las figuras que vamos a ver a continuación se corresponden a la representación los triángulos entrópicos (separados) obtenidos de los experimentos realizados.

La alteración de las probabilidades a priori de los datos de entradas debida a la reducción de los subconjuntos de datos citados anteriormente se ve claramente en los triángulos entrópicos separados (con la entropía de entrada y salida representadas) pues las primeras ya no están perfectamente alineadas. Este efecto tiene lugar sobre todo en la representación de la intersección con los de ISOLET y ruidosos pues al ser muy pocos y existir muchas confusiones con otros fonemas se modifican notablemente las entropías de entrada. Constatado entonces que existen muchas confusiones entre fonemas de M&N55 y otros, sospechamos que la mayoría de estas confusiones serán con otras consonantes (y no con vocales, salvo por problemas de alineamiento que no nos interesan para caracterizar las confusiones por similitud acústico-articulatoria). En este caso nos fijaremos en los resultados obtenidos de los coeficientes MFCC (clean y noisy) filtrados tanto paso bajo como paso alto.

Del mismo modo que en RHH, los triángulos que vamos a mostrar son de dos tipos:

- Con el color representando la precisión.
- Con el color representando la frecuencia de corte.

Aquí, además del filtrado paso alto o paso bajo añadimos el reconocimiento a partir de datos limpios o ruidosos.

4.2.2.1.1 Representación MFCC Clean paso bajo

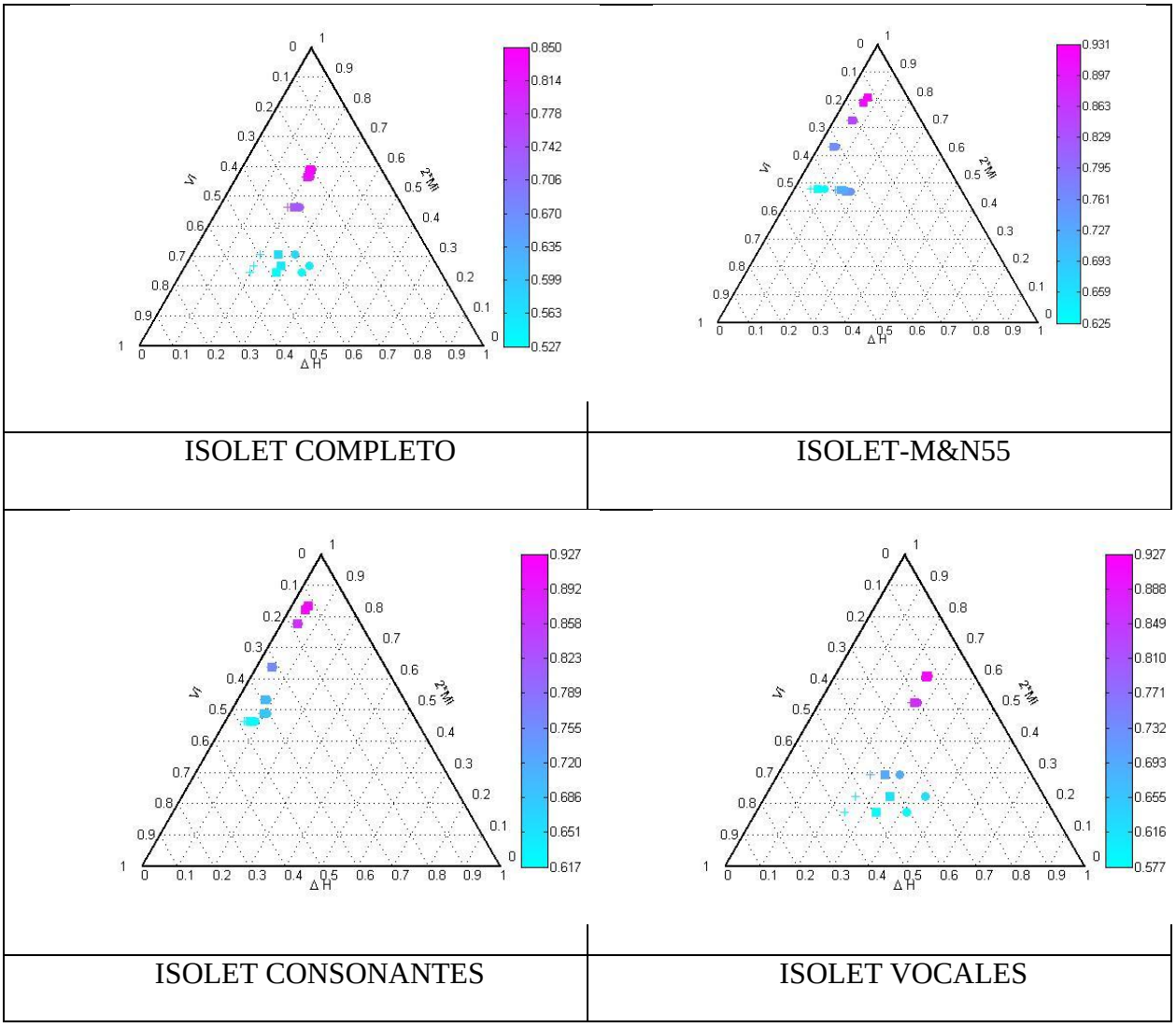

Figura 43: Triángulos entrópicos MFCC Clean paso bajo con representación de la precisión

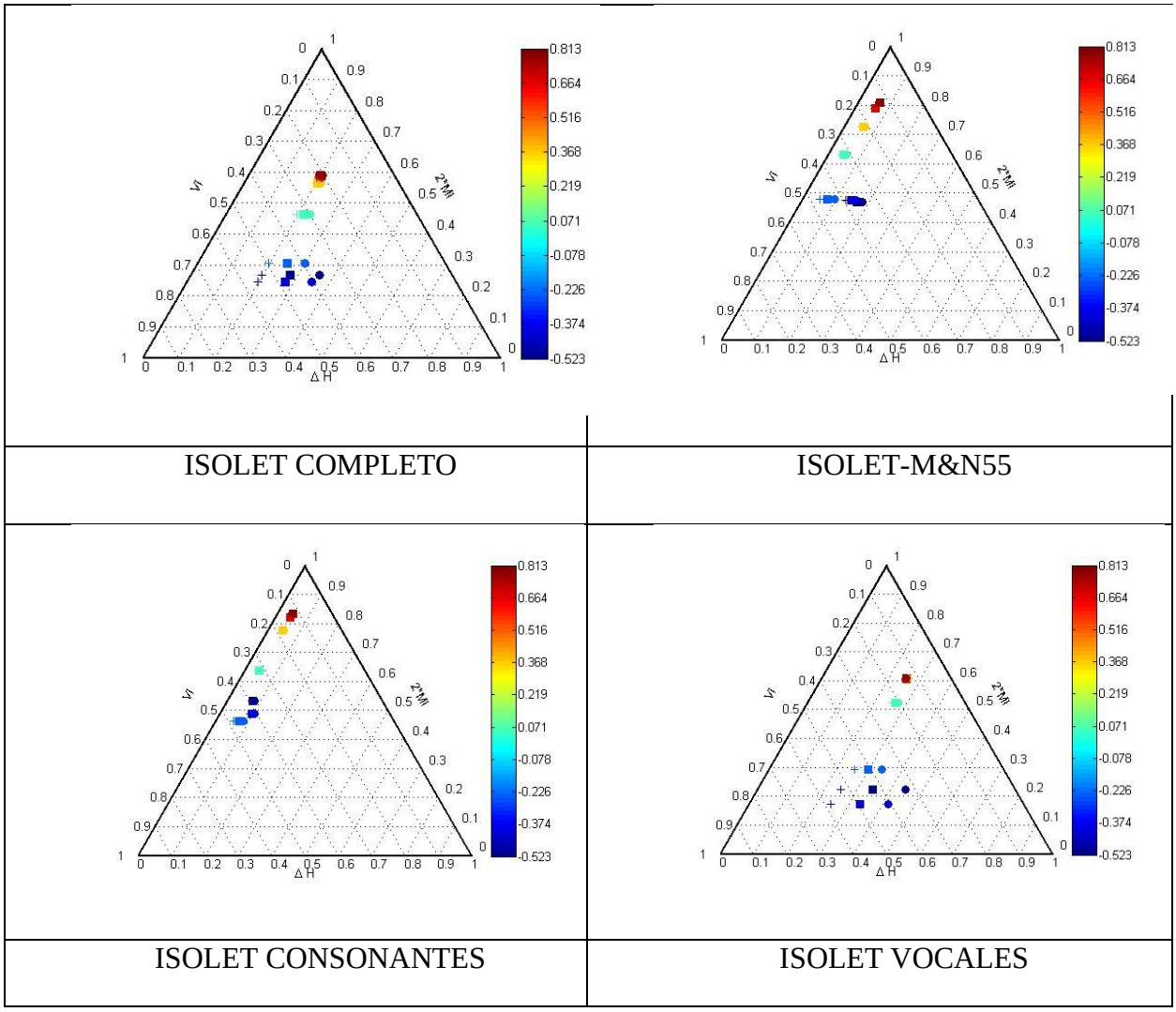

Figura 44: Triángulos entrópicos MFCC Clean paso bajo con representación de la frecuencia de corte.

Analizando las representaciones incluidas en las figuras 43 y 44 podemos observar:

Cuando representamos el conjunto de fonemas completo podemos percibir que las entradas están situadas hacia el centro del triángulo y la salida del reconocedor tiende a desplazarse hacia la derecha lo que nos indica que para conseguir mejores resultados se especializa o lo que es lo mismo, que la base de datos presenta gran desequilibrio entre las clases y eso introduce un sesgo importante hacia tomar decisiones basadas en las clases mayoritarias.

Al reducir los fonemas de entrada a los correspondientes a los experimentos de M&N55 podemos ver que los datos tienden a situarse más a la izquierda que en la representación completa lo que nos indica que el equilibrio entre las clases aquí es mayor. Como esto también se da en el caso de ISOLET CONSONANTES a diferencia del de ISOLET VOCALES, tenemos que concluir que es esta diferencia tan marcada entre la cantidad de muestras presentes de las vocales por un lado, y de las consonantes, por otro, lo que causa el problema del desequilibrio global mencionado anteriormente.

Además, podemos distinguir claramente que a la frecuencia de corte 200-400Hz sobre todo pero también la de 200-300 Hz su posición tiende a desplazarse hacia a la derecha quedando desalineado

del resto que vemos que tienen una progresión lineal. Estudiaremos este caso más detalladamente obteniendo los retículos de confusión correspondientes a esa frecuencia de corte.

En el caso de reducción a las consonantes vemos que el comportamiento del reconocedor es muy similar al caso de los M&N55, es decir, los datos tienden a colocarse a la izquierda del triángulo y hacia arriba lo que nos dice que el reconocedor se comporta de forma equilibrada para este conjunto de datos de entrada. Seguimos viendo como en los dos casos anteriores que a las frecuencias más bajas el reconocedor tiende a desplazarse hacia la derecha y hacia arriba para obtener mejores resultados mediante la especialización.

En cuanto a la representación de las vocales vemos que su comportamiento es muy distinto a las consonantes, podemos observar que en este caso los datos de entrada se encuentran más hacia el centro del triángulo y que la salidas a bajas frecuencias tienden a irse hacia la derecha buscando la especialización de manera mucho más pronunciada que en los caso anteriores. De los histogramas de la sección 4.1.3 ya sabemos que incluso dentro de las propias vocales hay desequilibrios importantes en el número de muestras con las que está representada cada clase.

Podemos concluir que la presencia de desequilibrios en la base de datos se debe sobre todo a las vocales y que esto hace que presente una tendencia hacia la especialización en ellas y que este efecto es más acusado cuando el ancho de banda es menor.

4.2.2.1.2 Representación MFCC Clean paso alto

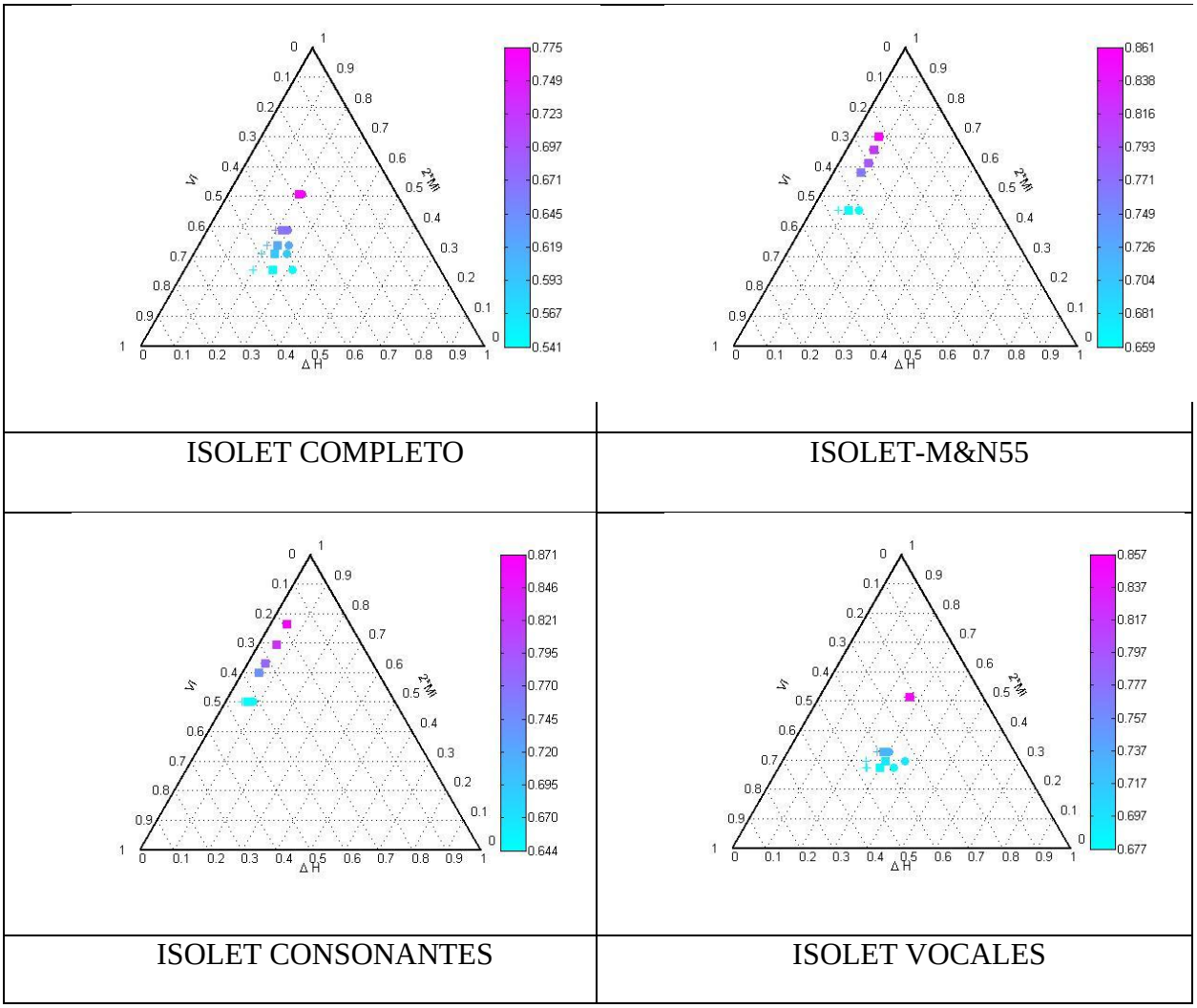

Figura 45: Triángulos entrópicos MFCC Clean paso alto con representación de la precisión

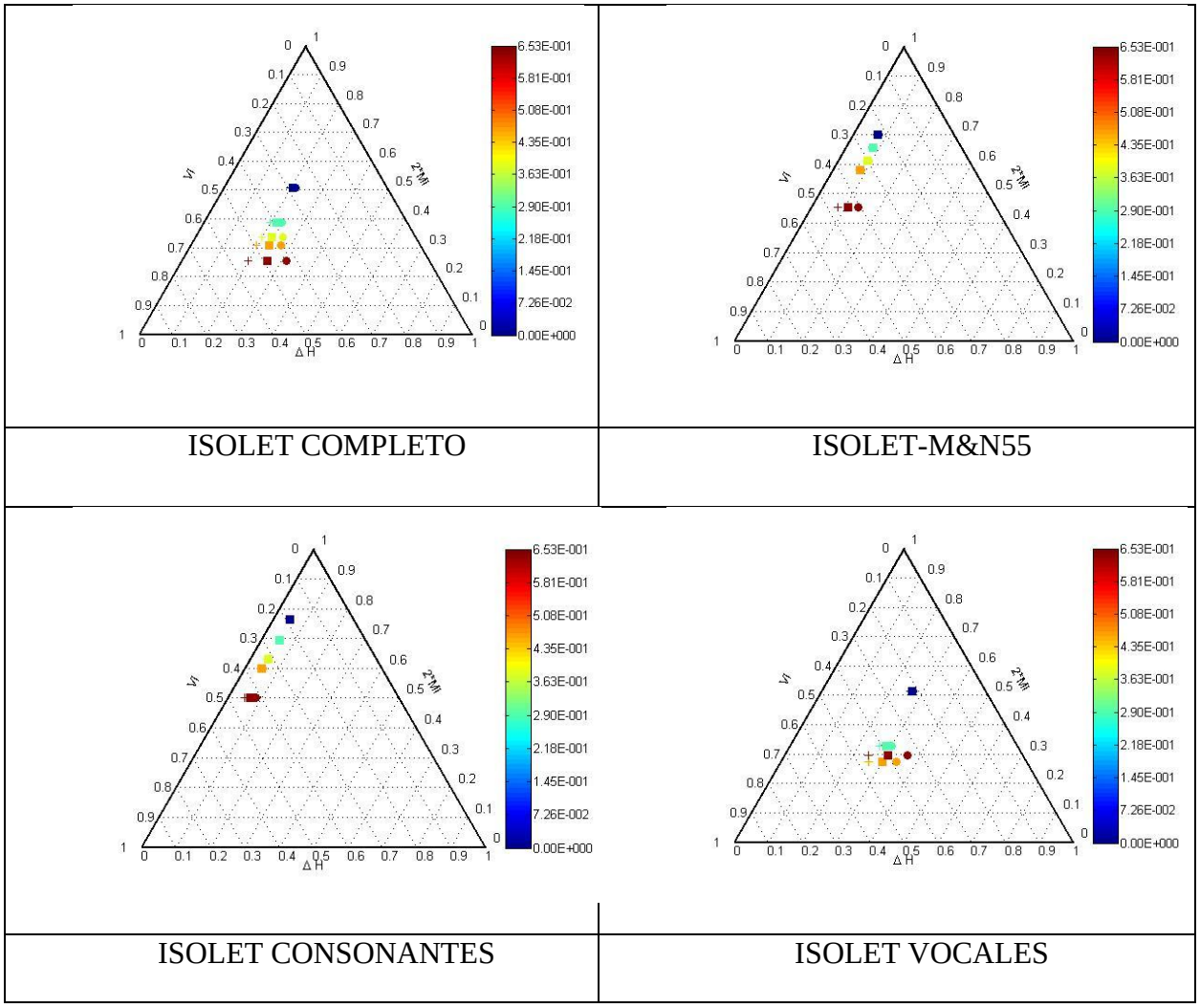

Figura 46: Triángulos entrópicos MFCC Clean paso alto con representación de la frecuencia de corte.

Analizando los Triángulos entrópicos representados en las figuras 45 y 46 observamos que:

1) Como en el caso paso bajo, el comportamiento para el conjunto de las consonantes tanto los alófonos comunes con M&N, como las consonantes de la base de datos ISOLET es bastante mejor que para el conjunto ISOLET total. Esto se cumple en todos los casos lo que quiere decir que el sistema modela muy bien las consonantes a pesar de estar infrarrepresentadas (en un principio debería haber más tramas vocálicas que consonánticas). Hay que tener en cuenta que los de las consonantes no tienen el silencio.

2) Es llamativo el mal comportamiento que tiene el clasificador para la vocales y la distancia tan grande observada entre la frecuencia inferior de 1000 Hz y las demás (2000, 2500, 3000 y 4500) quedándose los datos en la zona baja centro y hacia la derecha del triángulo. La gran distancia existente entre los datos de entrada y los de salida revela también, el intento de obtener mejoras en la precisión a base de especialización en las vocales mayoritarias.

3) Como ya hemos concluido anteriormente el clasificador obtiene mejores resultados para las consonantes que para las vocales desde el punto de vista del aprendizaje a pesar de que la precisión que presenta es, en ambos casos, muy parecida.

4.2.2.1.3 Representación MFCC Noisy paso bajo

A continuación añadimos un grado más de complejidad que no está presente en los experimentos RHH de M&N55 observando el comportamiento con ruido además del de filtrado.

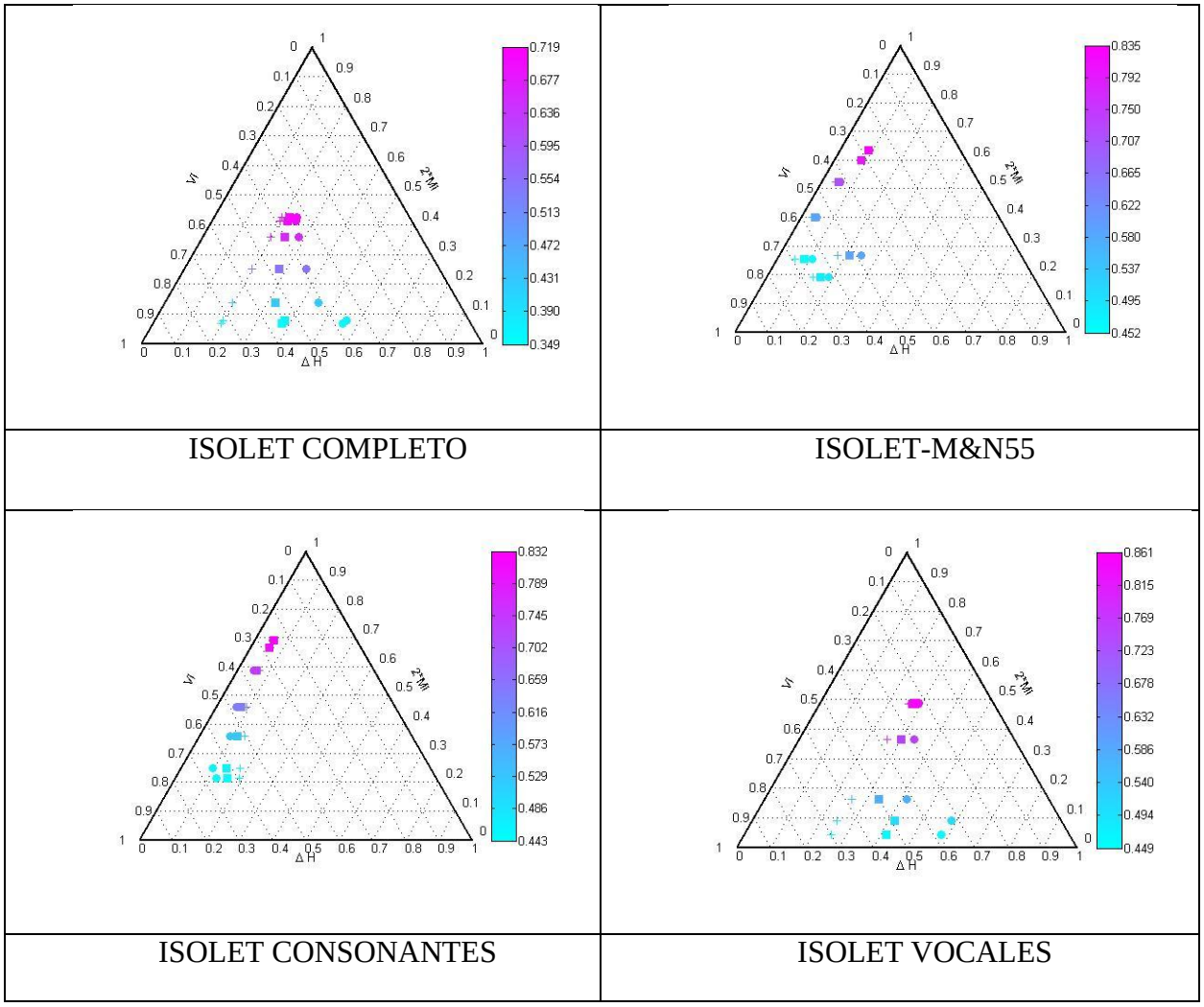

Figura 47: Triángulos entrópicos MFCC Noisy paso bajo con representación de la precisión

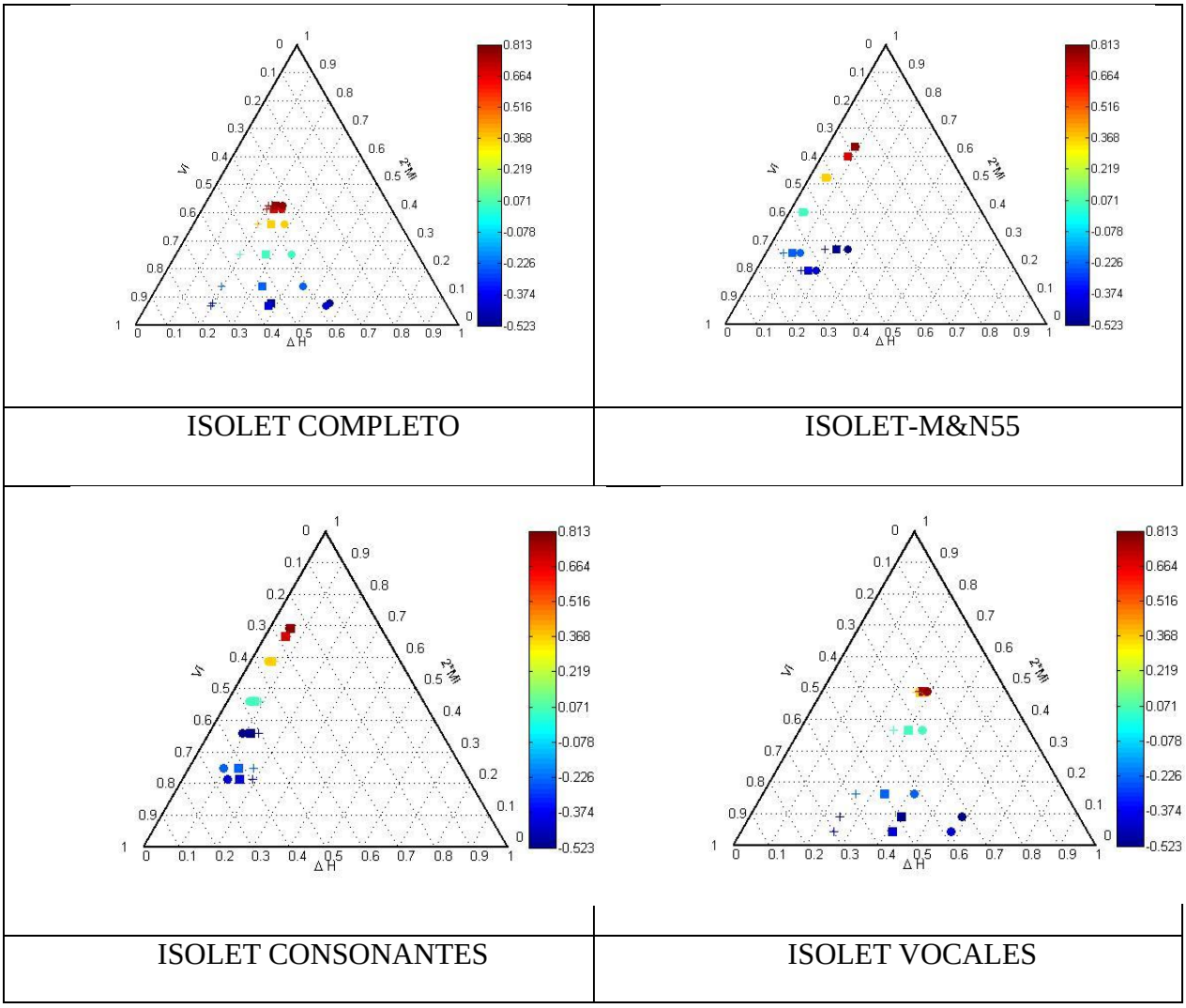

Figura 48: Triángulos entrópicos MFCC Noisy paso bajo con representación de la frecuencia de corte.

Analizando los triángulos entrópicos de las figuras 47 y 48 podemos concluir que:

1) El comportamiento es bastante mejor para las consonantes que para el conjunto total como ya hemos comentado en los casos anteriores. Esto se cumple en todos los casos lo que quiere decir que el sistema transmite mejor la información de las consonantes a pesar de estar infrarrepresentadas (en un principio debería haber más tramas vocálicas que consonánticas).

2) En todo caso podemos ver cómo el clasificador hace uso de una banda más ancha (en este caso añadir frecuencias más altas) mejorando la información mutua transmitida y reduciendo la entropía de la salida. Sin embargo, para los dos o tres últimos puntos (con frecuencias de corte más restringidas -300, 400 y 600 Hz-) los fonemas de ISOLET-M&N55 tienen un comportamiento extraño que no se observa en el resto de las configuraciones: podemos ver cómo la precisión no está directamente correlada con la información mutua pues se obtiene precisión a base de especialización con el desplazamiento hacia la derecha de los últimos dos puntos. Veremos en los retículos si podemos explicar esto observando las confusiones individuales. Este comportamiento también tiene lugar en los experimentos limpios aunque menos acusadamente.

3) Lo verdaderamente curioso es que se da en la intersección ISOLET-M&N55 de esta forma tan marcada pero también en el caso de sólo consonantes se ve el efecto curioso de obtener un mejor comportamiento para la frecuencia de corte superior de 300 Hz en comparación con las de 400 y 600 Hz que se plasma en la inversión del orden de los colores que observamos en el triángulo con la representación de la frecuencia de corte. También en el caso de las vocales, observamos una inversión de los colores esta vez sólo entre las frecuencias superiores de 300 y 400 Hz. Esto nos hace sospechar que con estos anchos de banda tan exiguos y combinados con el ruido, el aprendizaje es verdaderamente difícil y que no hay gran diferencia entre 100, 200 e incluso 500 Hz de ancho de banda.

4.2.2.1.4 Representación MFCC Noisy paso alto

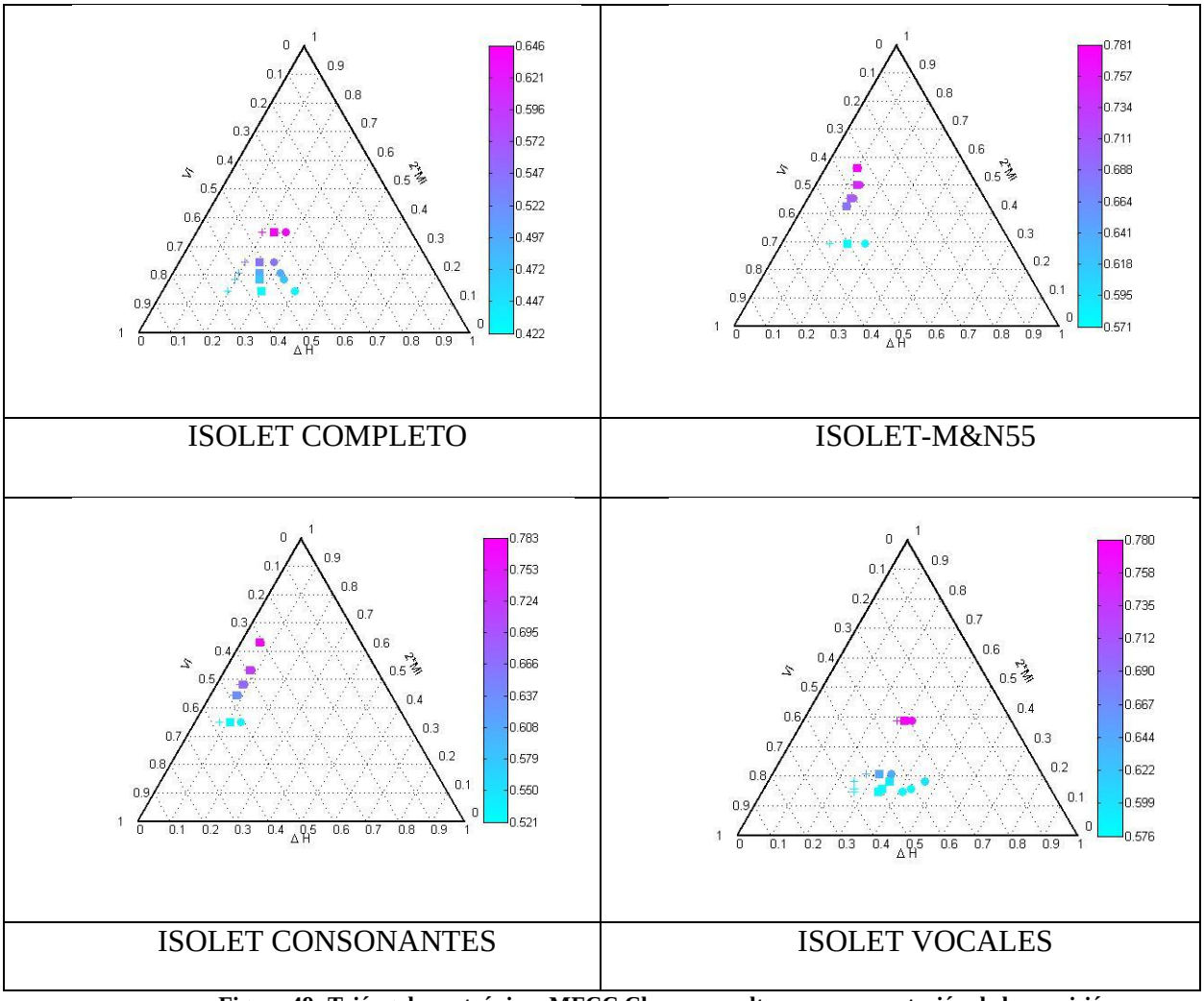

Figura 49: Triángulos entrópicos MFCC Clean paso alto con representación de la precisión

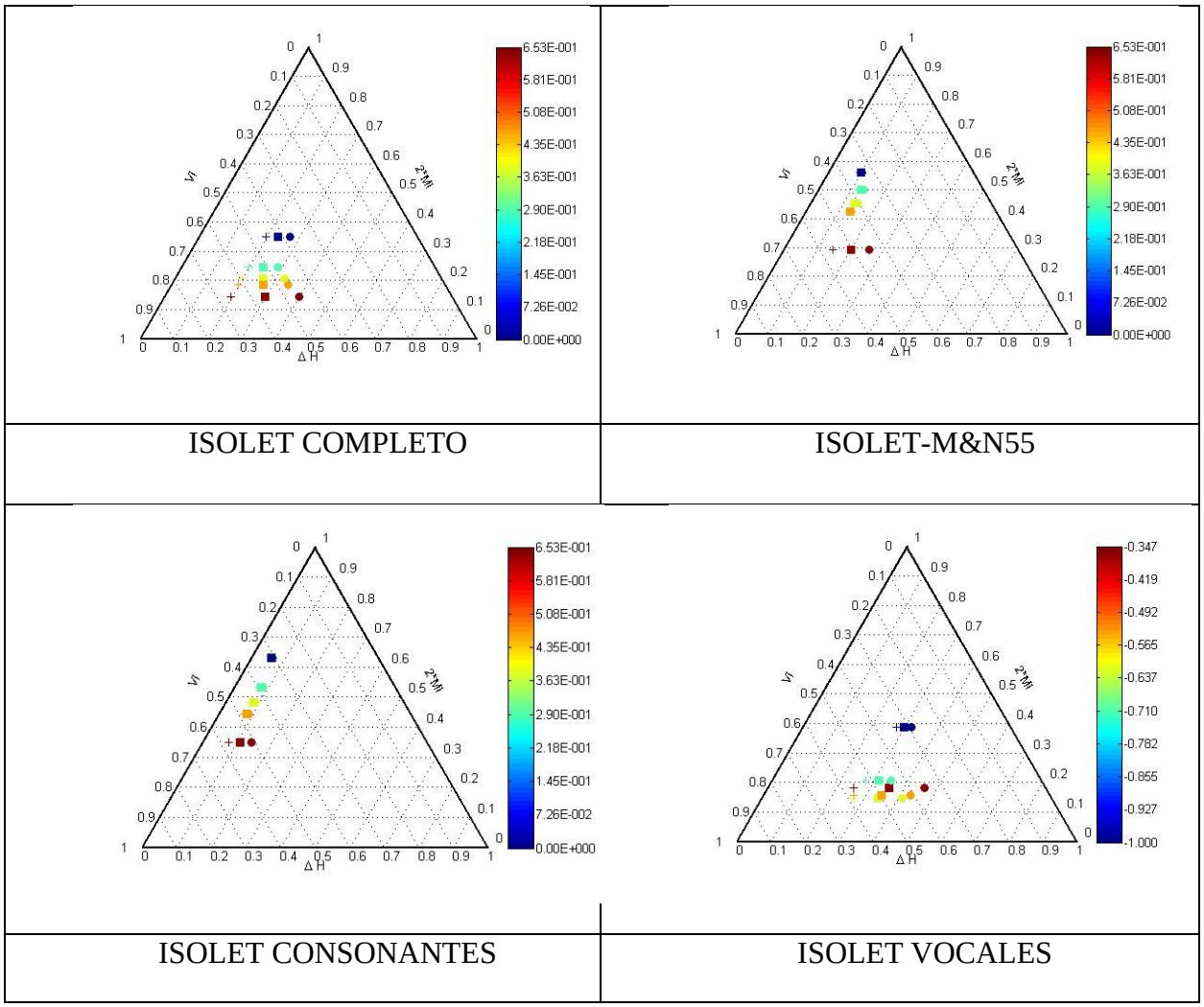

Figura 50: Triángulos entrópicos MFCC Noisy paso alto con representación de la frecuencia de corte.

Tras analizar los triángulos representados en las figuras 49 y 50 podemos observar que:

1) Se sigue observando el mismo comportamiento general en la comparación de las consonantes con las vocales que ya ha sido comentado en los casos anteriores. También volvemos a observar la gran distancia existente entre el ancho de banda superior y los demás en el caso de las vocales que ya habíamos observado en los filtrados paso alto sin ruido así como numerosas inversiones cuando los anchos de banda disminuyen lo que nos da una idea de la dificultad de modelar las vocales en estas situaciones.

2) En todo caso podemos ver cómo el clasificador hace uso de una banda más ancha (en este caso de añadir bajas frecuencias) mejorando la información mutua transmitida y reduciendo la entropía de la salida.

4.2.2.2 Retículos de confusión

Una vez estudiados los triángulos entrópicos pasamos a analizar de forma más pormenorizada las frecuencias con mayor interés representadas anteriormente.

Para este estudio obtendremos los retículos de confusión e interpretaremos los mismos.

Para el caso del experimento RAH vamos a representar los retículos correspondientes a los subconjuntos ISOLET-M&N55, ISOLET CONSONANTES e ISOLET VOCALES con las siguientes frecuencias de corte:

200-300Hz, 200-400Hz, 200-600Hz, 200-5000Hz. 1000-5000Hz, 2000-5000Hz y 3000-5000Hz.

Incluiremos aquí sólo las correspondientes a los coeficientes MFCC sin ruido ya que los ruidosos no contribuyen de forma importante a la discusión posterior y los presentamos en un apéndice.

4.2.2.2.1 Frecuencia 200-300Hz.

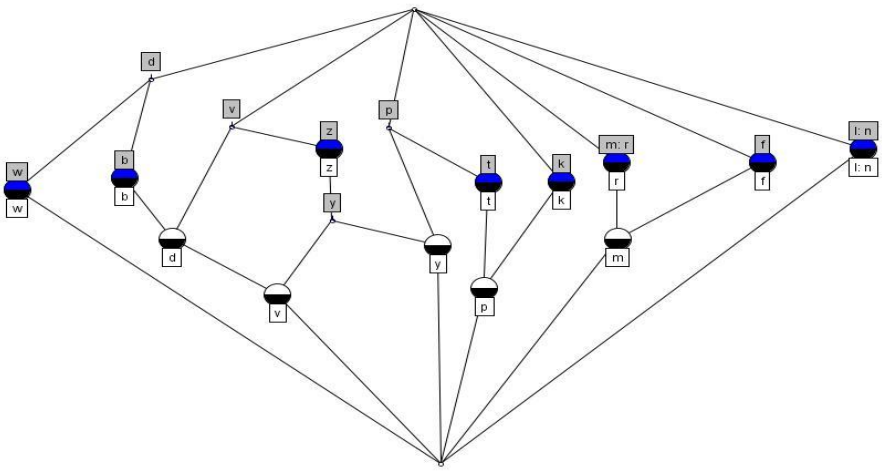

Figura 51: Retículo de confusión del subconjunto ISOLET CONSONANTES MFCC Clean con φ= -0.626738 y 19 conceptos.

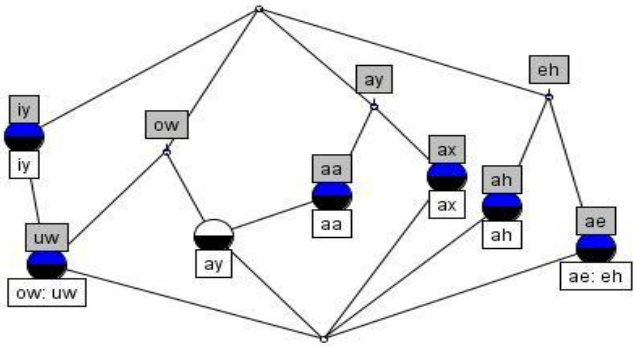

Figura 52: Retículo de confusión del subconjunto ISOLET VOCALES MFCC Clean con φ= -0.392199 y 12 conceptos.

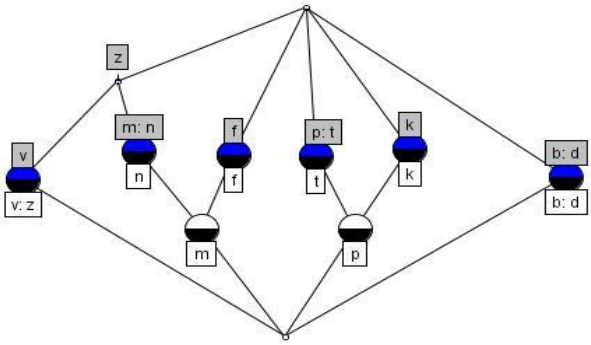

Figura 53: Retículo de confusión del subconjunto ISOLET-M&N55 MFCC Clean con φ= -2.67675 y 11conceptos.

En el caso del retículo de confusión representado para el experimento ISOLET reducido solo al conjunto de sus consonantes (figura 51) podemos ver claramente que para la frecuencia seleccionada las confusiones se agrupan de la siguiente manera:

1) A una frecuencia tan restringida y baja se producen muchas más confusiones y menos previsibles que en frecuencias de corte mucho menos restrictivas.

2) /l/ y /n/ es el primero de los sub-retículos adjuntos que comparten el punto de articulación y la sonoridad.

3) El segundo está formado por /m/ fonema nasal que se confunde con los fonemas /f/ y /r/.

4) Además vemos como el conjunto formado por las oclusivas sordas /p/, /t/, /k/ aparece continuamente en nuestros análisis RAH así que podemos concluir que esta característica sí que está bien capturada por los modelos automáticos y además es bastante robusta a las distorsiones debidos a las reducciones de ancho de banda.

5) En este caso podemos ver a la izquierda de la figura como los fonemas pertenecientes a las letras del denominado E-SET /b/, /d/, /v/, /z/ se confunden entre sí. Todos ellos son sonoros aunque los dos primeros son oclusivos y los dos últimos fricativos. Destacamos que el fonema oclusivo sonoro que falta, /g/, no aparece en ninguno de estos retículos por estar correctamente clasificado.

6) Cabe destacar la aparición del fonema /w/ que se conoce como semi-vocal confundiéndose con /d/ que pertenece a las oclusivas alveolares.

7) Si pasamos a observar para esta misma frecuencia el retículo de confusión representado para el conjunto de ISOLET VOCALES (figura 52) observamos que sólo hay dos sub-retículos adjuntos: el formado por /ah/, /ae/ y /eh/ y el resto de las vocales que están relacionadas entre sí. Es lógico dado que a la frecuencia analizada es difícil conseguir distinguirlas.

Además recordamos del análisis de los triángulos entrópicos que este reconocedor trabaja mucho mejor para las consonantes que para los vocales.

En la figura 53 que corresponde al retículo de confusión del subconjunto ISOLET-M&N55 vemos que las confusiones se agrupan en:

1) /b/ y /d/, define un conjunto de fonemas que solo se confunden entre ellos mismos quedando separados del resto, es decir, forman un sub-retículo adjunto al del resto de los fonemas. De estos conjuntos /b/ y /d/ tienen en común que son fonemas oclusivos sonoros y pertenecen al conjunto de las denominadas parejas de confusiones comunes E-SET [21].

2) Además volvemos a ver el conjunto /p/, /k/, /t/ de consonantes oclusivas sordas.

3) Por otro lado vemos cómo /m/ fonema correspondiente a las consonantes nasales se confunde con /n/ su pareja de confusión más común y a su vez con /f/ y también con /z/ ambas fricativas.

4) Por su parte /n/ se confundirá con /z/ y está con su pareja más común la /v/ consonante fricativa como la /z/.

4.2.2.2.2 Frecuencia 200-400Hz.

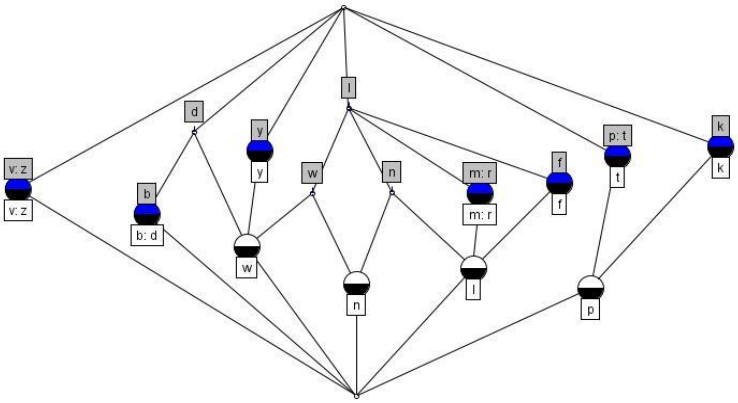

Figura 54: Retículo de confusión del subconjunto ISOLET CONSONANTES MFCC Clean con phi -0.992567 y 17 conceptos.

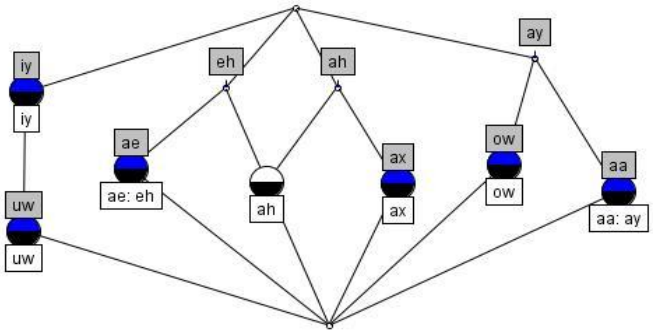

Figura 55: Retículo de confusión del subconjunto ISOLET VOCALES MFCC Clean con φ= -0,325508 y 12 conceptos.

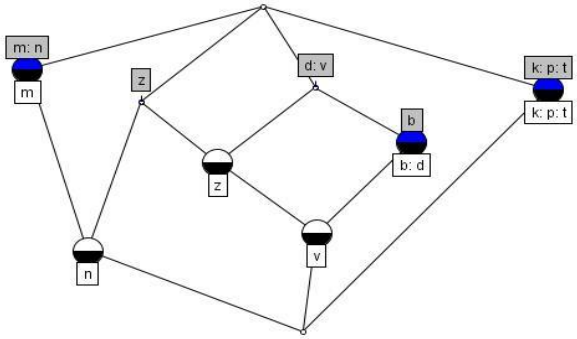

Figura 56: Retículo de confusión del subconjunto ISOLET-M&N55 MFCC Clean con phi 1,024072 y 10 conceptos.

En el caso del retículo de confusión representado para el experimento ISOLET CONSONANTES (figura 54) así como el de ISOLET-M&N55 (Figura 56) podemos ver que son bastante similares al de las frecuencias anteriores (200-300 Hz):

1) /v/ y /z/ define el primer sub-retículos adjunto. De estos conjuntos /v/ y /z/ tienen en común que son fonemas fricativos y pertenecen al E-SET [21].

2) Vemos como el conjunto formado por /p/, /t/, /k/ vuelve a aparecer confundiéndose la /p/ con /k/ y con /t/.

/l/ se confunde con /m/, /r/ y /f/, viendo que ninguna de ellas salvo /l/ y /r/ tienen nada en común.

4) Vuelven a juntarse algunos fonemas pertenecientes al E-SET /b/, /d/ aunque se siguen mezclando con /w/, /y/.

En la figura 56 que corresponde al retículo de confusión del subconjunto ISOLET-M&N55 vemos que las confusiones se agrupan:

1) Una vez más las oclusivas sonoras /k/, /p/, /t/ independientes todas confundidas con todas.

2) Vemos como parte de las consonantes de conjunto E-SET se confunden entre ellas en la parte central de la figura /b/, /d/, /v/, /z/.

3) Notamos que los fonemas nasales /m/ y /n/ vuelven a aparecer en este retículo como pareja de confusiones comunes relacionándose con la /z/ a través de la /n/ como en el caso anterior.

4) Sin embargo, el retículo de ISOLET VOCALES es bastante diferente del anterior (figura 55): observamos tres sub-retículos adjuntos /uw/ y /iy/ por una parte, /ow/, /aa/ y /ay/ por otra y el resto.

4.2.2.2.3 Frecuencia 200-600Hz.

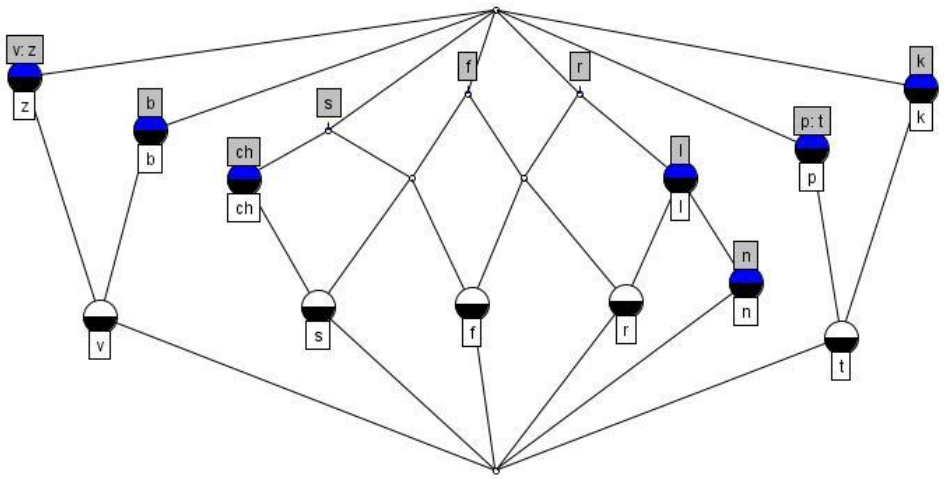

Figura 57: Retículo de confusión del subconjunto ISOLET CONSONANTES MFCC Clean con φ=--0.747149y 19 conceptos.

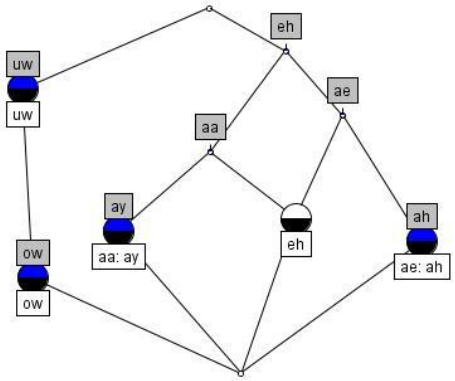

Figura 58: Retículo de confusión del subconjunto ISOLET VOCALES MFCC Clean con φ= -0,164087 y 10 conceptos.

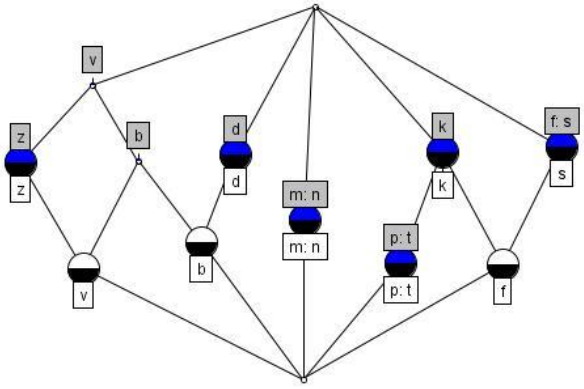

Figura 59: Retículo de confusión del subconjunto ISOLET-M&N55 MFCC con φ= 0,086399 y 13 conceptos.

A pesar de que en este caso el ancho de banda sigue siendo muy restrictivo empezamos a ver algunas novedades en este análisis sobre todo en el retículo ISOLET-M&N55 en el que empezamos a ver una estructura de confusiones más acorde con la teoría acústico-fonética.

En el caso del retículo de confusión representado para el experimento ISOLET CONSONANTES (figura 57) destacamos la aparición de los fonemas /s/ y /ch/ que hasta ahora se había mantenido como clasificado sin errores y la desaparición de los fonemas /m/, /d/, /w/ y /y/ por estar correctamente clasificadas. Sin embargo, sí aparece la /n/ que se confunde con la consonante también sonora /l/.

Vemos como el conjunto formado por /p/, /t/, /k/ vuelve a aparecer confundiéndose la /t/ con /k/ y con /p/. Por otra parte, /v/, /z/ y /b/ vuelven a confundirse entre ellas sabiendo que son fonemas pertenecientes al E-SET.

Si pasamos a observar para esta misma frecuencia el retículo de confusión para ISOLET VOCALES (figura 58) observamos que han desaparecido varias vocales que ya han dejado de confundirse con otras. Observamos dos sub-retículos adjuntos: /uw/ y /ow/ por una parte y el resto de las vocales representadas. Recordaremos que el triángulo entrópico correspondiente mostraba que este clasificador estaba menos especializado que los dos anteriores.

En la figura 59 que corresponde al retículo de confusión ISOLET-M&N55 encontramos una estructura más parecida a la de las encontrados en RHH. También el triángulo entrópico señalaba ya en este punto una tendencia hacia la izquierda alineada con el resto de los casos de filtrado paso bajo menos restrictivos que separan este caso de los dos anteriores analizados:

El sub-retículo nasal aparece claramente con /m/ y /n/, también el de las oclusivas sordas /p/, /t/, /k/ (aunque ligadas a /s/ y /f/) y como a la izquierda de la figura se agrupan confundiéndose entre sí algunos de los fonemas pertenecientes al conjunto E-SET. En este caso /v/, /z/, /b/ y /d/.

4.2.2.2.4 Frecuencia 200-5000Hz.

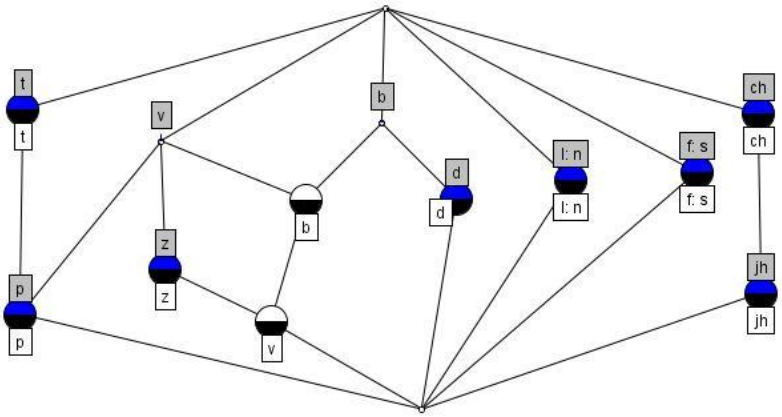

Figura 60: Retículo de confusión del subconjunto ISOLET CONSONANTES MFCC Clean con φ= 0,668614 y 14 conceptos.

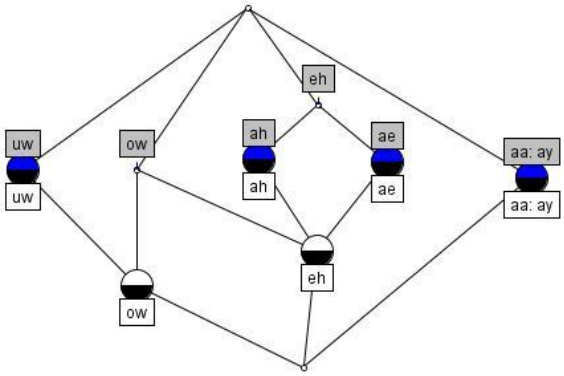

Figura 61: Retículo de confusión del subconjunto las ISOLET VOCALES MFCC Clean con φ= 1,962855 y 10 conceptos.

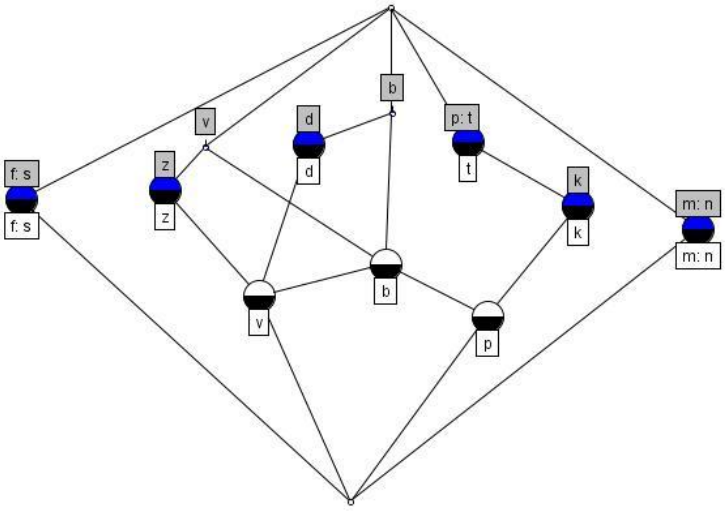

Figura 62: Retículo de confusión del subconjunto ISOLET-M&N55 MFCC Clean con φ= 2,411295y 13 conceptos.

En esta frecuencia tenemos el caso para el cual se puede considerar que los conjuntos de datos de entrada no están filtrados.

En el caso del retículo de confusión representado para el experimento ISOLET CONSONANTES (figura 60) podemos ver que para la frecuencia seleccionada las confusiones se agrupan de la siguiente manera:

1) /ch/ y /jh/, /f/ y /s/, /l/ y /n/ forman tres sub-retículos adjuntos. Son confusiones que no hemos encontrado en RHH y que por tanto son errores propios del modelado automático.

2) El cuarto sub-retículo está formado por el resto de los fonemas representados. Observamos mucha más presencia que hasta el momento desde la parte central de la figura hacia la izquierda como todos los fonemas que se confunden entre sí corresponden al E-SET. Sin embargo, no aparecen (por estar perfectamente clasificados) /m/, /k/ y /g/.

3) En cuanto a las vocales (figura 61) observamos que la estructura de confusiones es muy parecida a la que observamos para la banda 200-600 Hz así que parece que no existen mejoras importantes para la captura de las vocales al incluir frecuencias superiores.

En la figura 62 (ISOLET-M&N55) observamos confusiones muy parecidas a las de RHH:

1) El sub-retículo de las nasales, /m/ y /n/, el de las fricativas sonoras /f/ y /s/ (labio-dental y alveolar resp.), las oclusivas sordas /p/, /t/ y /k/ enlazadas con las sonoras /b/ y /d/ a través del nexo entre la /p/ y la /b/ y con las fricativas /v/ y /z/ a través de /b/ y /d/.

4.2.2.2.5 Frecuencia 1000-5000Hz.

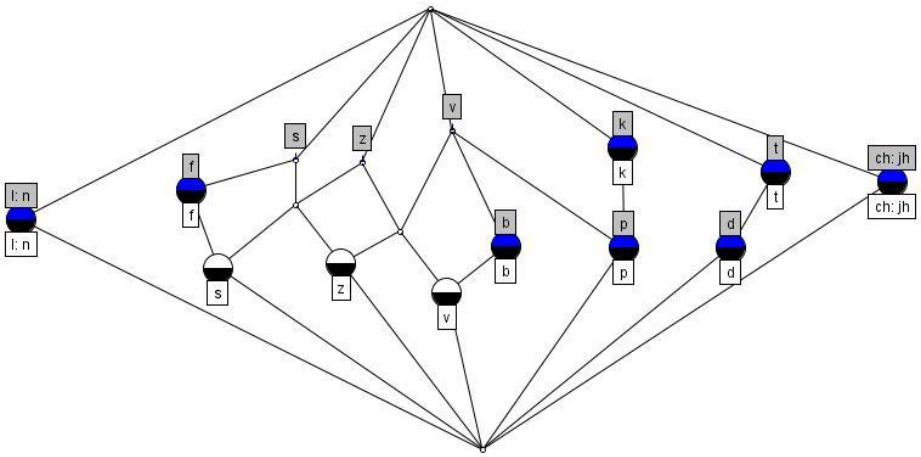

Figura 63: Retículo de confusión del subconjunto ISOLET CONSONANTES MFCC Clean con φ=- 0,431464 y 18 conceptos.

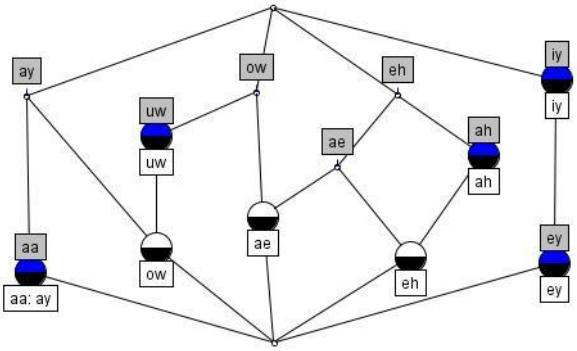

Figura 64: Retículo de confusión del subconjunto las ISOLET VOCALES MFCC Clean con φ= 1,13314 y 14 conceptos.

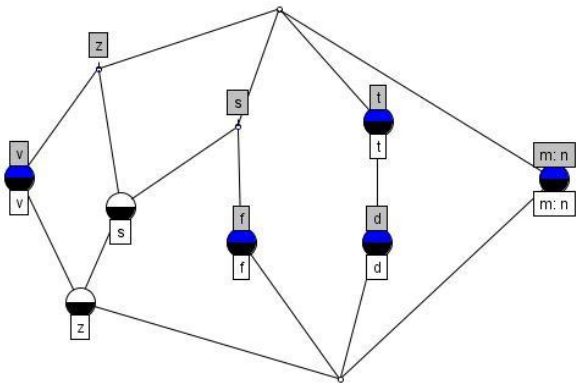

Figura 65: Retículo de confusión del subconjunto ISOLET-M&N55 MFCC Clean con φ= 0,647846 y 11 conceptos.

El primero de los retículos que analizamos con un borrado de frecuencias bajas (paso alto) ISOLET CONSONATES (figura 63) presenta unas confusiones parecidas al de la banda completa. Recordemos que los triángulos entrópicos no preveían desviaciones importantes más allá de un deterioro de las prestaciones. Resaltamos las diferencias:

1) /ch/ y /jh/, /l/ y /n/ son sub-retículos adjuntos ya presentes en el caso anterior. /d/ y /t/ sin embargo, es un sub-retículo que aparece nuevo y que denota una confusión en la sonoridad puesto que ambos son fonemas oclusivos alveolares. Sin embargo, no aparecen (por estar perfectamente clasificados) /m/ y /g/ (antes también teníamos aquí el fonema /k/).

2) El sub-retículo /f/ y /s/ que aparecía independiente en el caso anterior ahora presentan enlaces con el resto de las fricativas /z/ y /v/ y esta última (labio-dental) a su vez con los fonemas labiales oclusivos /b/ (sonoro) y /p/ (sordo).

3) Observamos de nuevo como es constante la confusión entre los fonemas /p/, y /k/, esta vez sin /t/ aunque en este caso /p/ también se confunde con /v/.

Si pasamos a observar para esta misma frecuencia el retículo de confusión representado para el conjunto de las vocales de la base de datos ISOLET (figura 64) observamos un nuevo sub-retículo formado por /iy/ y /ey/ que antes aparecían sin confusiones, además de las confusiones habituales que ya veníamos observando.

En la figura 65 (ISOLET-M&N55) observamos cómo se han desligado completamente las oclusivas de las fricativas a diferencia de lo que observábamos en la Figura 62 de tal forma aparecen las nasales (/m/ y /n/) por un lado, las fricativas /v/,/z/, que a su vez se confunde con la pareja de confusión ya conocida y definida en [21] /f/ y /s/ por otro y las oclusivas /t/ y /d/ que comparten punto y modo de articulación y difieren sólo en la sonoridad, como hemos comentado ya. Es notable que el resto de las oclusivas ya no aparecen por estar correctamente clasificadas.

4.2.2.2.6 Frecuencia 2000-5000Hz.

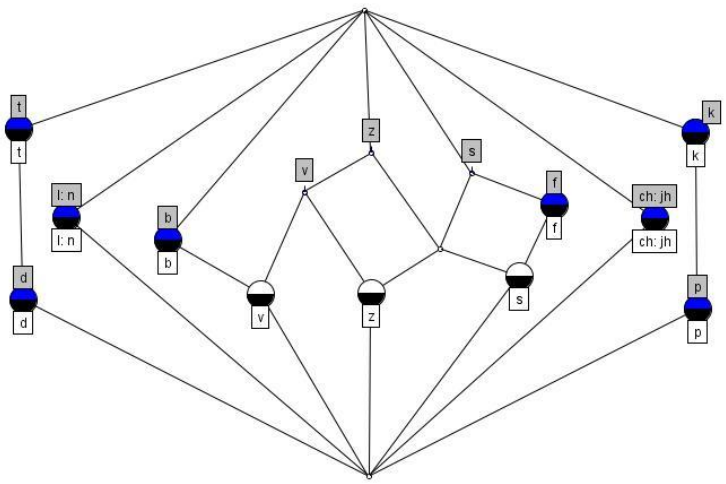

Figura 66: Retículo de confusión del subconjunto ISOLET CONSONANTES MFCC Clean con φ=-0.068744 y 18 conceptos.

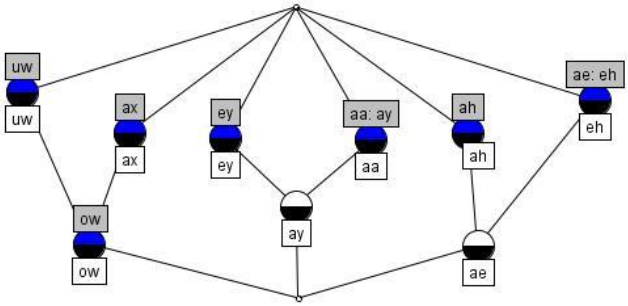

Figura 67: Retículo de confusión del subconjunto ISOLET VOCALES MFCC Clean con φ= 0,459171 y 11 conceptos.

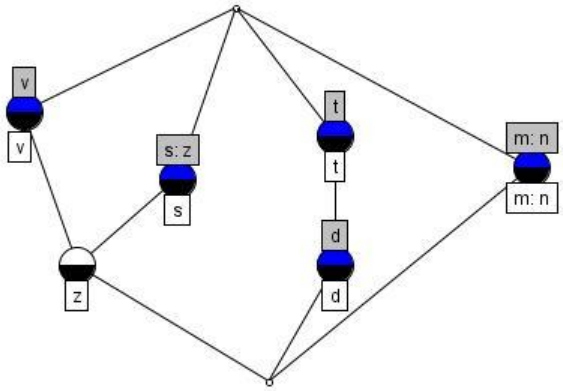

Figura 68: Retículo de confusión del subconjunto ISOLET-M&N55 MFCC Clean con φ= 0,218961 y 8 conceptos.

Las confusiones en ISOLET CONSONANTES (Figura 66) para este rango de frecuencias es prácticamente idéntico al del rango anterior (Figura 63) siendo la única diferencia la ausencia del nexo /v/ - /p/. Es llamativa la poca repercusión que tiene en la estructura de los errores la ausencia de la banda de 1000 a 2000 Hz.

Desde el punto de vista de las vocales (figura 67) vemos algo más de estructura con tres sub-retículos adjuntos: /uw/, /ow/ y /ax/; /ey/, /ay/ y /aa/; /ae/, /ah/, /eh/.

En la figura 65 (ISOLET-M&N55) observamos cómo se han desligado completamente las oclusivas de las fricativas a diferencia de lo que observábamos en la Figura 62 de tal forma aparecen las nasales (/m/ y /n/) por un lado, las fricativas /v/,/z/, que a su vez se confunde con la pareja de confusión ya conocida y definida en [21] /f/ y /s/ por otro y las oclusivas /t/ y /d/ que comparten punto y modo de articulación y difieren sólo en la sonoridad, como hemos comentado ya. Es notable que el resto de las oclusivas ya no aparecen por estar correctamente clasificadas.

En la figura 68 (ISOLET-M&N55) volvemos a observar (como en la Figura 65) los tres sub-retículos: fricativas, oclusivas y nasales donde lo único diferente es la ausencia de /f/.

4.2.2.2.7 Frecuencia 3000-5000Hz.

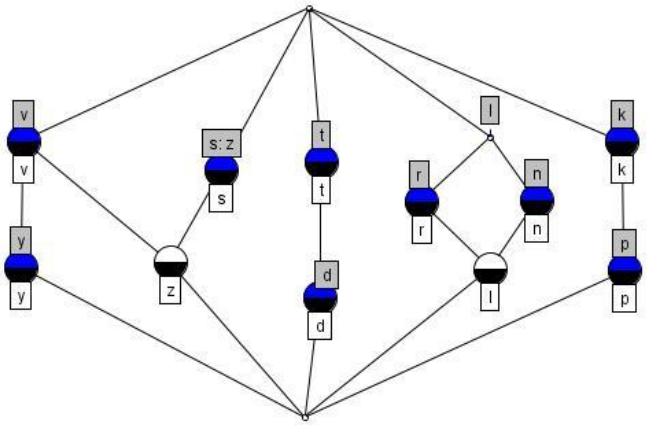

Figura 69: Retículo de confusión del subconjunto ISOLET CONSONANTES MFCC Clean con φ=- - 0,416725 y 14 conceptos.

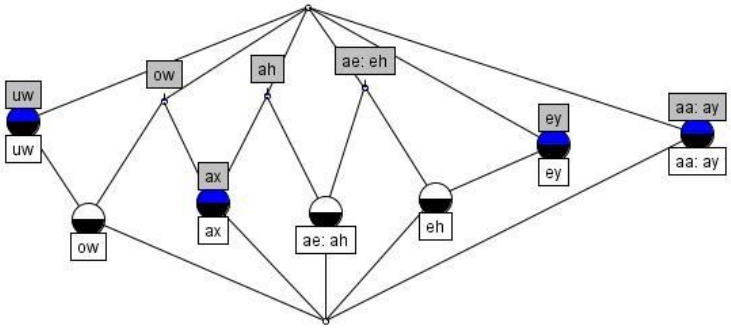

Figura 70: Retículo de confusión del subconjunto ISOLET VOCALES MFCC Clean con φ= 0.047668 y 12 conceptos.

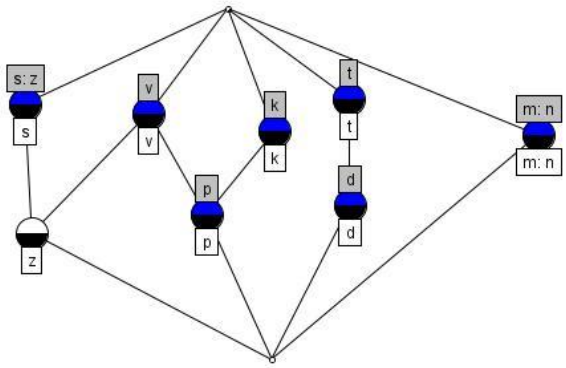

Figura 71: Retículo de confusión del subconjunto ISOLET-M&N55 MFCC Clean con φ= 0,591854 y 17 conceptos.

En esta banda de frecuencias mucho más restrictiva nos encontramos, sin embargo, que la estructura general de los errores se mantiene. Sin embargo, aparecen algunos fonemas añadidos que vamos a comentar a continuación:

1) En el caso del retículo de confusión para ISOLET CONSONANTES (figura 69) podemos ver que además de los sub-retículos /t/-/d/ y /p/-/k/ que veníamos observando, el sub-retículo /l/-/n/ aparece aumentado con /r/, consonante líquida como /l/ que comparten el punto de articulación.

2) A la izquierda de la figura tendremos las confusiones de /z/ con /s/ y /v/ que ya habíamos observado a las que se unen las confusiones de /y/ con /v/. El fonema /y/ sólo lo habíamos visto en la banda de 200-400 (Figura 54) confundido con la también semivocal /w/ aunque en este caso aparece confundido con /v/ curiosamente.

3) No encontramos nada especialmente reseñable en ISOLET VOCALES (figura 70) donde vemos que hemos vuelto a perder la estructura que observábamos en la Figura 67.

4) El retículo para ISOLET-M&N55 (figura 71) es prácticamente idéntico al anterior (Figura 68) salvo por el nexo entre /p/ y /v/ que, aunque con puntos de articulación cercanos no comparten ni la sonoridad ni el modo de articulación.

4.2.3 COMPARATIVA DE RESULTADOS.

Una vez presentados y analizados los resultados obtenidos de los experimentos pasamos a comparar el comportamiento del experimento RHH con el RAH. La comparativa la haremos con los alófonos comunes de ambos experimentos.

1) Observando los resultados y el análisis de los mismos vemos que los experimentos de RHH tienen mejores resultados, como esperábamos, ya que el reconocedor utilizado es el propio ser humano.

2) En los triángulos de entropía vemos que en el caso RAH los resultados se desplazan a la derecha del triángulo en comparación con los RHH, dicho desplazamiento esta ocasionado por el tratamiento de los datos explicado anteriormente 4.1.3.

3) En cuanto a los retículos de confusión vemos que el comportamiento de ambos reconocedores difiere de los canales presentados por M&N 55:

Vemos como consonantes nasales como /m/ y /n/ no sólo se confunden entre sí, sino que lo hacen con consonantes oclusivas como la /k/.

4) También observamos que aunque las confusiones tanto en el experimento RHH como el RAH tienen un comportamiento parecido, destacamos que aunque en el RHH aparecen siempre /p/ y /t/ como una confusión común en el caso RAH se les une /k/, viendo como en casi todas las frecuencias y figuras analizadas aparece como un conjunto de consonantes /p/, /t/,/k/ .

Dicha aparición puede proporcionarnos información útil sobre las características articulatorias.

5 CONCLUSIONES Y LÍNEAS FUTURAS.

Tras la comparación de los resultados de los experimentos RHH con los RAH concluimos que el reconocedor RHH sigue siendo mejor que el RAH aunque éste da resultados notables para las consonantes.

Gracias a las herramientas presentadas a lo largo de este documento hemos podido estudiar tanto la calidad del clasificador como sus confusiones.

Dichas herramientas nos han presentado formas innovadoras de estudiar la calidad y las confusiones de un reconocedor obteniendo a su vez, información muy útil sobre el comportamiento del mismo que nos permitirán abrir futuras líneas de mejora y de estudio de los reconocedores de habla actuales.

El hecho de no fijarnos solo en la precisión como medida de calidad del reconocedor nos permitirá estudiar confusiones importantes que en otros casos pasarían desapercibidas.

Gracias al análisis formal de conceptos hemos podido estudiar de forma clara las confusiones del reconocedor viendo la importancia de las características articulatorias divididas en canales por M&N55: sonoridad, nasalidad, punto de articulación, fricción y duración.

De este estudio podemos concluir la importancia de las características articulatorias que deben cumplir los reconocedores para asemejarse lo máximo posible a RHH. Cuanto más estudiemos el comportamiento de RHH y a su vez implementemos mejoras en los RAH actuales los resultados obtenidos podrán parecerse más al RHH y por tanto ser mejores.

El hecho de analizar la calidad de un reconocedor con los triángulos entrópicos (ET) nos abre las puertas a un nuevo concepto de entender la calidad del reconocedor que merece ser estudiado en más detalle y desarrollarlo más puesto que con la utilización de esta herramienta podemos tener una visión mucho más objetiva y real del comportamiento del clasificador.

6 PRESUPUESTO

Todo proyecto siempre tiene un apartado de presupuesto en el que se debe analizar el tiempo dedicado a él, los medios utilizados, los costes que le han producido al proyectando, y luego aplicarle el margen de beneficio que se estime oportuno que se quiere ganar en el proyecto realizado.

Para conocer el esfuerzo del proyectando, además de la etapa de ejecución del mismo hay que considerar ciertas tareas que también fue necesario realizar, y que se han considerado para calcular su duración:

- Estudios de antecedentes y documentación: Consiste en buscar información relacionada con el proyecto, antecedentes, búsqueda de documentación útil, etc. Con esta tarea se comienza el proyecto, pero es una actividad continua durante todo el desarrollo de éste.
- Tiempo dedicado al aprendizaje de las herramientas usadas: (Linux, matlab, scripts)
- Realización de los experimentos.

- Preparación, análisis y presentación de los resultados obtenidos.
- Redacción de la memoria.

Para calcular el presupuesto total de este trabajo, vamos a desglosarlo según los costes ocasionados:

6.1 COSTES DEL PERSONAL

Estos costes serán la suma de los costes del proyectando y de la persona encargada de la dirección del proyecto.

- En este caso la persona encargada de llevar a cabo los experimentos y las conclusiones obtenidas de ellos es el proyectando.

Apellidos, nombre	categoría profesional	Dedicación (persona/mes)	Coste (persona/mes)	Coste (euros)
González Martín, Sira	ingeniero técnico	12	2693,54	32322,48

- Costes de dirección del proyecto.

Estos costes son los asociados a la dirección de este proyecto, supervisión y revisión del mismo.

Descripción	Coste/euros	% dedicación	dedicación (meses)	Coste imputable
dirección	20000	10	12	20000

6.2 COSTES DERIVADOS DEL EQUIPAMIENTO UTILIZADO.

Dos ordenadores personales del proyectando, cuyo coste se detalla a continuación:

Descripción	Coste/euros	% dedicación	dedicación (meses)	periodo depreciación	Coste imputable
Intel Core Duo	500	100	12	60	100
AMD Phenom™ II X4	500	100	12	60	100
total					200

Intel Core Dúo: Usado para realizar la programación del presente proyecto y la generación de la memoria correspondiente además de la realización de los experimentos y análisis de los mismos

AMD Phenom™ II X4: Usado para realizar la programación del presente proyecto y la generación de la memoria correspondiente además de la realización de los experimentos y análisis de los mismos

6.3 COSTES DE FUNCIONAMIENTO

Otros costes relacionados con la creación del proyecto son los siguientes:

Microsoft Office 2007: El Office es necesario para poder realizar el documento final de la memoria del proyecto.

MATLAB: este programa ha sido usado para programar diferentes funciones que han hecho posible la realización de los experimentos y una vez realizados han sido la base para poder preparar y analizar los datos obtenidos. Requiere de licencia para su uso.

Tarifa ADSL: Una conexión a Internet ha sido necesaria para poder descargar tanto la base de datos como el reconocedor además de para obtener información y documentación relacionado con el proyecto. El coste de la tarifa es de 36 € y se ha utilizado durante 10 meses.

Sistema operativo Unix (Linux): el reconocedor de habla usado en este proyecto se ha desarrollado para usarse en entornos Unix por ello, para la realización del mismo se han usado distribuciones gratuitas de debian, más concretamente Ubuntu.

Sistema operativo Windows : es el sistema operativo en el cual se ha llevado a cabo la memoria, no produce costes adicionales al venir como sistema operativo en los dos ordenadores usados.

Base de datos ISOLET: es la base de datos de nuestros experimentos, no genera costes al poderse descargar de forma gratuita.

Entorno experimentación ISOLET TESBED: es el reconocedor que hemos usado para nuestros experimentos, tampoco genera coste alguno al poderse descargar de forma gratuita.

Descripción	Coste imputable
Microsoft Office 2007	140
MATLAB	70
Tarifa ADSL	360
Sistema operativo Unix (Linux):	0

Base de datos ISOLET	0
Entorno experimentación ISOLET TESBED	0
Total	570

6.4 RESUMEN DE LOS COSTES

Para calcular los costes totales se han tenido en cuenta una tasa de costes indirectos del 30%.

Si sumamos todos estos costes obtendremos el total del presupuesto de este proyecto.

presupuesto costes totales	presupuesto costes
Personal	32322,48
Equipamiento utilizado	200
Subcontratas	
Costes funcionamiento	570
Costes indirectos	15927,744
Costes dirección	20000
Total	**69020,224**

7 REFERENCIAS

[1] *Miller, George A.; Nicely, Patricia E. An analysis of perceptual confusions among some English consonants. The Journal of the Acoustical Society of America, 2005, vol. 27, no 2, p. 338-352.*

[2] *Definición de fricación. Disponible en: : http://rephip.unr.edu.ar/bitstream/handle/2133/1367/5..FON%C9TICA%20y%20FONOLOGIA.pdf;jsessionid=97B85440EDC0D09C4CE61DE6AE3C1795?sequence=6*

[3] *Definición consonante africada, disponible en : http://es.wikipedia.org/wiki/Consonante_africada*

[4] *McGill W. J , Psychometrika 19, 97-116 (1954)*

[5] *Definición RAH disponible en (http://es.wikipedia.org/wiki/Reconocimiento_del_habla)*

[6] *Ilustración. Disponible en: (http://liceu.uab.cat/~joaquim/speech_technology/tecnol_parla/recognition/speech_recognition/reconocimiento.html)*

[7] *Peña, Vicente, et al. Contribuciones al reconocimiento robusto de habla. 2007.*

[8] *L.E. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes", 1972 Inequalities, vol. 3, pp. 1-8.*

[9] *Ljolie, A.; Ephraim, Y.; Rabiner, L. R. Estimation of hidden Markov model parameters by minimizing empirical error rate. En Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on. IEEE, 1990. p. 709-712.*

[10] *Boulard H and Morgan NHybrid HMM/ANN Systems for Speech Recognition: Overview and New Research Directions. Technical report, IDIAP, Martigny, Switzerland. Intl. Comp. Science Institute, Berkeley, CA. UC Berkeley, Berkeley, CA, 1998.*

[11] *Trentin, E. and M Gori A survey of hybrid ANN/HMM models for automatic speech recognition .ITC-irst (Centro per la Ricerca Scientixca e Tecnologica), V. Sommarive, 18-Povo, Trento, Italy and Universitadi Firenze, V. S. Marta, 3 - Firenze, Italy Dipartimento di Ingegneria dell'Informazione, Universita di Siena, V. Roma, 56 - Siena, Italy Received 27 April 1999; accepted 15 April 2000*

[12] *Scharenborg, Odette. Reaching over the gap: A review of efforts to link human and automatic speech recognition research. Speech Communication, 2007, vol. 49, no 5, p. 336-347.*

[13] *Arie Ben-David (2007),A lot of randomness is hiding in accuracy*

[14] *Valverde-Albacete, Francisco J.; Pélaez-Moreno, Carmen. 100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox. PloS one, 2014, vol. 9, no 1, p. e84217.*

[15] *Mejía-Navarrete, David, et al. Feature extraction assessment for an acoustic-event classification task using the entropy triangle.*

[16] *Loizou, Philipos C.; Spanias, Andreas S. High-performance alphabet recognition. Speech and Audio Processing, IEEE Transactions on, 1996, vol. 4, no 6, p. 430-445.*

[17] *Definición precision y recall http://en.wikipedia.org/wiki/Precision_and_recall*

[18] *Valverde-Albacete, Francisco José; Carrilo-de-Albornoz, Jorge; Pélaez-Moreno, Carmen. A proposal for new evaluation metrics and result visualization technique for sentiment analysis tasks. En Information Access Evaluation. Multilinguality, Multimodality, and Visualization. Springer Berlin Heidelberg, 2013. p. 41-52.*

[19] *Cohen, J.A., 1960. Coefficient of agreement for nominal scales. Educational and Psychological Measurement, 37–4]*

[20] *Valverde-Albacete, Francisco J.; Pélaez-Moreno, Carmen. Two information-theoretic tools to assess the performance of multi-class classifiers.Pattern Recognition Letters, 2010, vol. 31, no 12, p. 1665-1671.*

[21] *Pélaez-Moreno, C.; García-Moral, A. I.; Valverde-Albacete, F. J. Analyzing phonetic confusions using formal concept analysis. The Journal of the Acoustical Society of America, 2010, vol. 128, no 3, p. 1377-1390.*

[22] *Ganter, Bernhard; Wille Rudolf. Formal concept analysis: mathematical foundations. Springer-Verlag New York, Inc., 1997.*

[23] *Definición de diagrama Hasse, disponible en: http://es.wikipedia.org/wiki/Diagrama_de_Hasse),*

[24] *Definición base de datos ISOLET disponible en : http://archive.ics.uci.edu/ml/datasets/ISOLET)*

[25] *Obtención de herramientas isolet testbed. Disponible en :http://www1.icsi.berkeley.edu/Speech/papers/gelbart-ms/hybrid-testbed/*

[26] *Obtención paquete SprachCore, reconocedor de habla automático. Disponible en:http://www1.icsi.berkeley.edu/~dpwe/projects/sprach/sprachcore.html*

[27] *Obtener base datos ISOLET (base datos sin modificar) Disponible en: http://www.icsi.berkeley.edu/ftp/pub/speech/papers/eurospeech05-onset/isolet/*

[28] *Obtención sistema operativo para los ordenadores. Disponible en http://www.ubuntu.com/*

[29] *Herramienta representación FCA : CONEXP , disponible en http://conexp.sourceforge.net/*

8 ANEXO: RETÍCULOS DE CONFUSIÓN CON BORRADOS DE FRECUENCIA Y RUIDO AMBIENTE

Añadimos aquí el análisis correspondiente a los retículos de confusión para la base de datos ISOLET contaminada con ruido que corresponde con los triángulos entrópicos presentados en las secciones 4.2.2.1.3 y 4.2.2.1.4. Dado que los experimentos de M&N55 no tienen correlato con estos y que los triángulos entrópicos indican que la tendencia de los errores es similar a la encontrada en la base de datos limpia hemos considerado que se desviaba del hilo argumental del proyecto y de ahí que se presente como un anexo.

8.1 FRECUENCIA 200-300Hz.

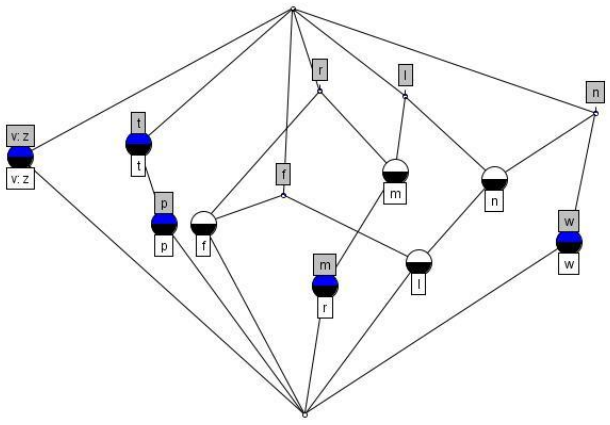

Figura 72: Retículo de confusión del subconjunto ISOLET CONSONANTES MFCC Noisy con φ= -1,944925y 15 conceptos.

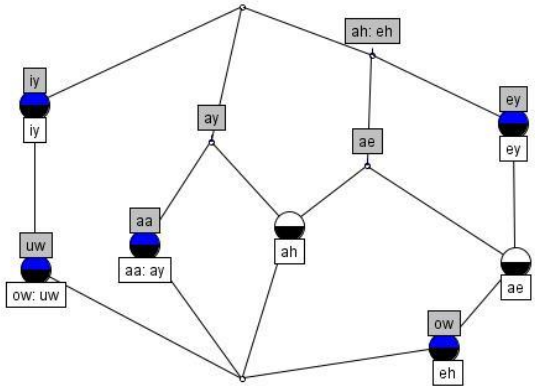

Figura 73: Retículo de confusión del subconjunto las ISOLET VOCALES MFCC Noisy con φ= -0,168368 y 12 conceptos.

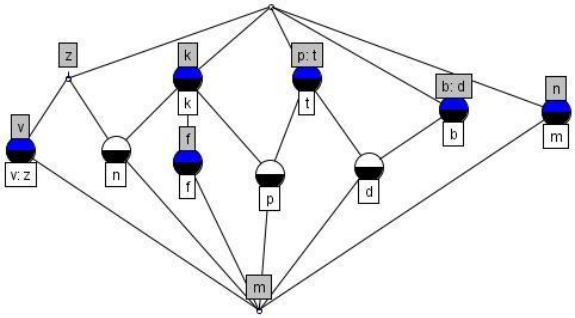

Figura 74: Retículo de confusión del subconjunto ISOLET-M&N55 MFCC Noisy con φ= -0,645063 y 12 conceptos.

8.2 FRECUENCIA 200-400Hz.

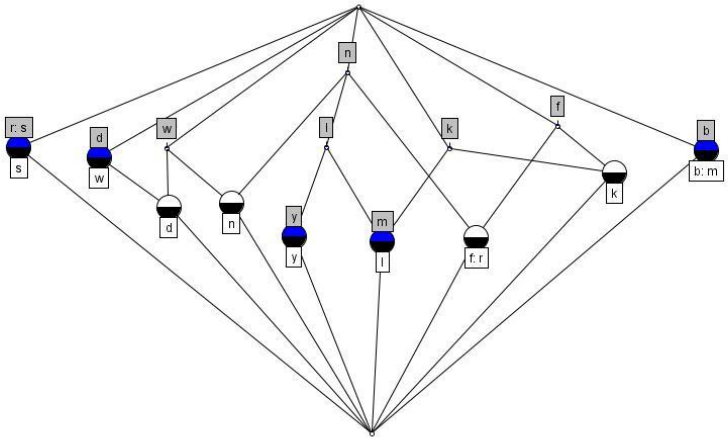

Figura 75: Retículo de confusión del subconjunto ISOLET CONSONANTES MFCC Noisy con phi -1,390295 y 16 conceptos.

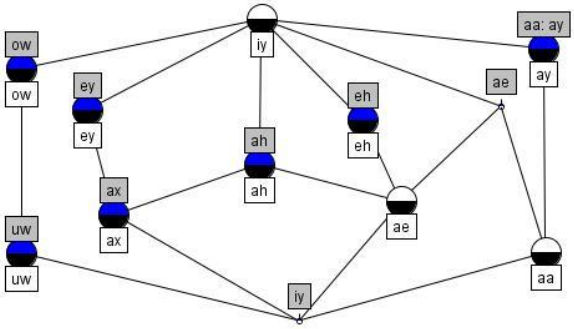

Figura 76: Retículo de confusión del subconjunto ISOLET VOCALES MFCC Noisy con φ= --0.842650 y 12 conceptos.

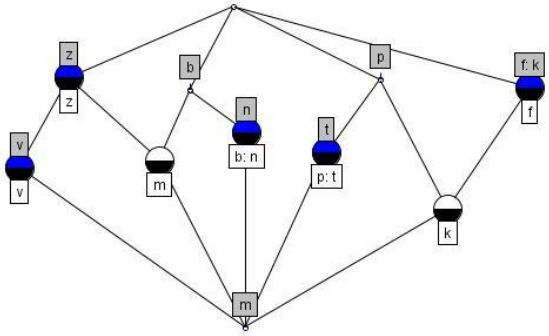

Figura 77: Retículo de confusión del subconjunto ISOLET-M&N55 MFCC Noisy con phi -0.716009 y 11 conceptos.

8.3 FRECUENCIA 200-600Hz.

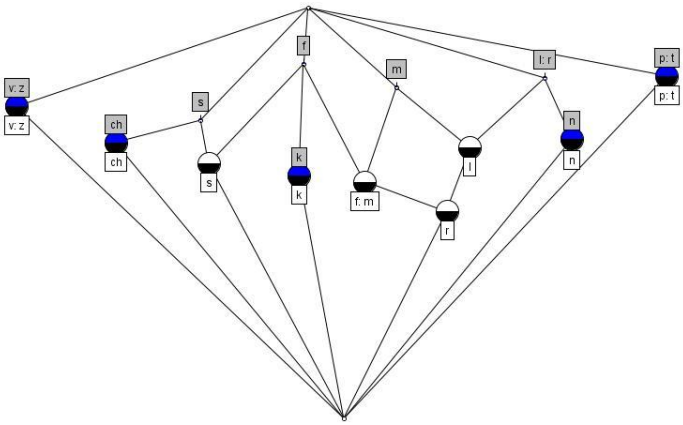

Figura 78: Retículo de confusión del subconjunto ISOLET CONSONANTES MFCC Noisy con phi -1,885011 y 15 conceptos.

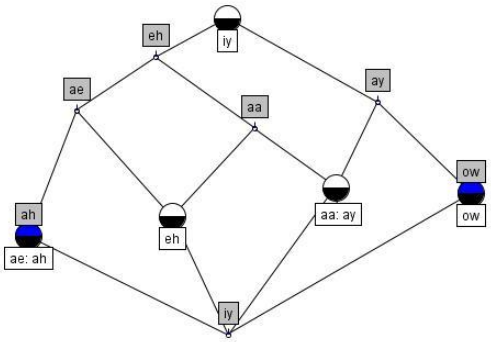

Figura 79: Retículo de confusión del subconjunto ISOLET VOCALES MFCC Noisy con φ= --0.774532 y 10 conceptos.

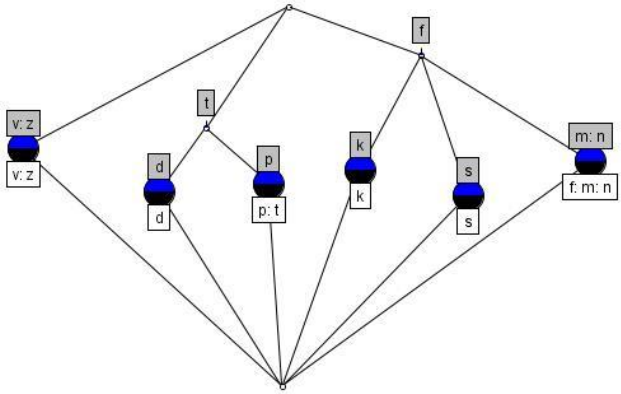

Figura 80: Retículo de confusión del subconjunto ISOLET-M&N55 MFCC Noisy con φ= -0,676845 y 10 conceptos.

8.4 FRECUENCIA 200-5000Hz.

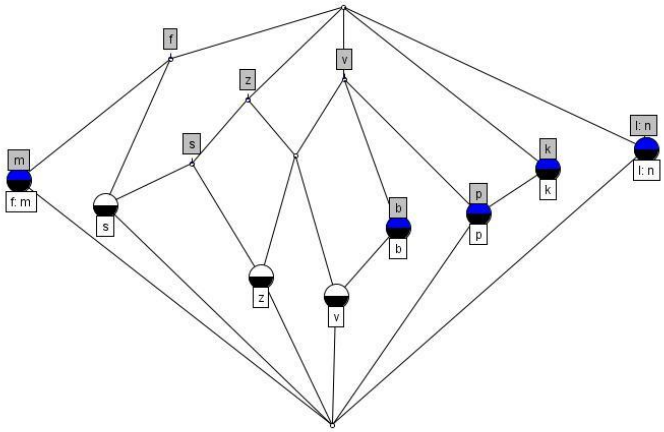

Figura 81: Retículo de confusión del subconjunto ISOLET CONSONANTES MFCC Noisy con φ= - 0,275992 y 15 conceptos.

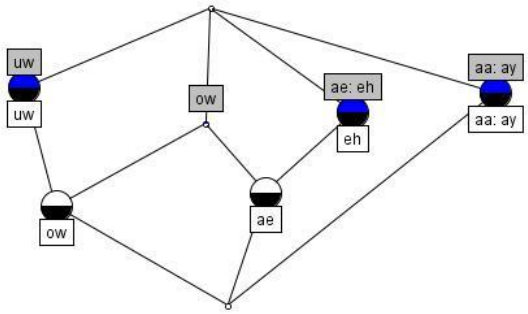

Figura 82: Retículo de confusión del subconjunto ISOLET VOCALES MFCC Noisy con φ= 0.066854 y 8 conceptos.

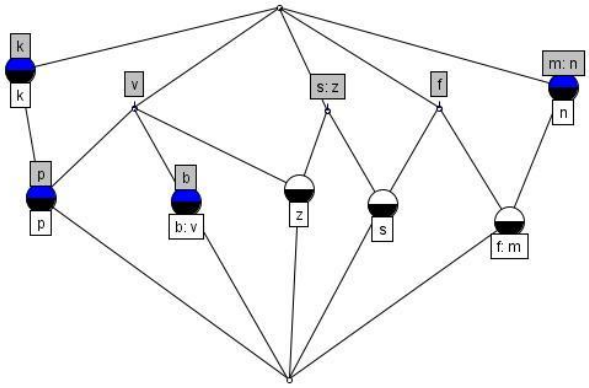

Figura 83. Retículo de confusión del subconjunto ISOLET MFCC Noisy con φ= 0,704643 y 12 conceptos.

8.5 FRECUENCIA 1000-5000Hz.

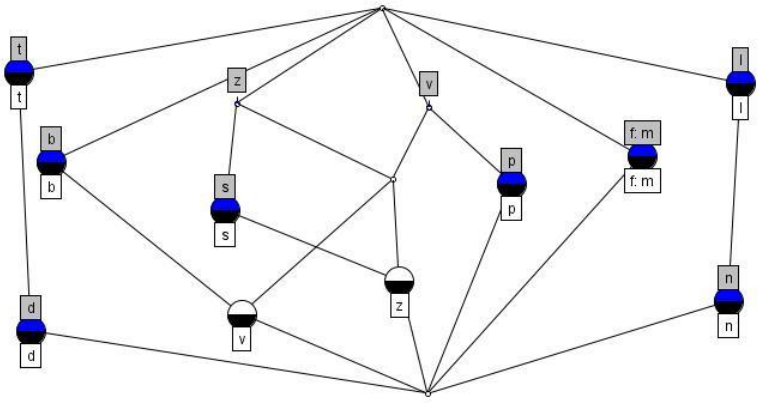

Figura 84: Retículo de confusión del subconjunto ISOLET CONSONANTES MFCC Noisy con phi -0,71333 y 15 conceptos.

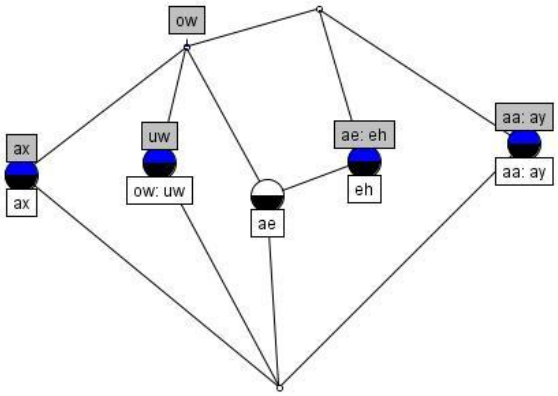

Figura 85: Retículo de confusión del subconjunto ISOLET VOCALES MFCC Noisy con φ= 0.031643 y 8 conceptos.

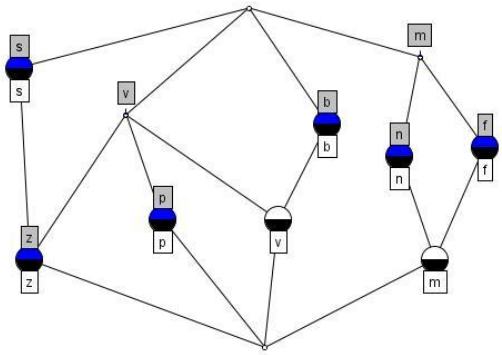

Figura 86: Retículo de confusión del subconjunto ISOLET MFCC-M&N55 Noisy con φ= -0.129686 y 12 conceptos.

8.6 FRECUENCIA 2000-5000Hz.

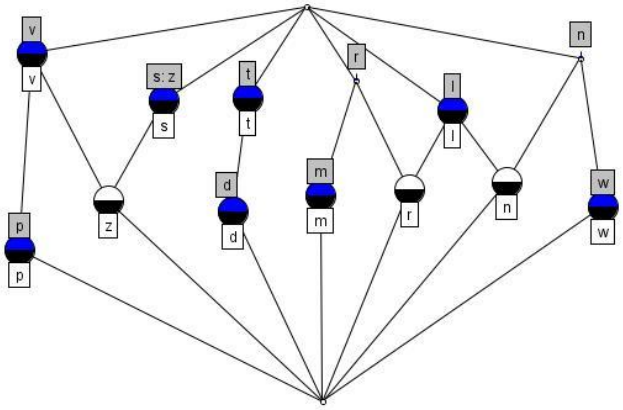

Figura 87: Retículo de confusión del subconjunto ISOLET CONSONANTES MFCC Noisy con phi -0,9523588 y 15 conceptos.

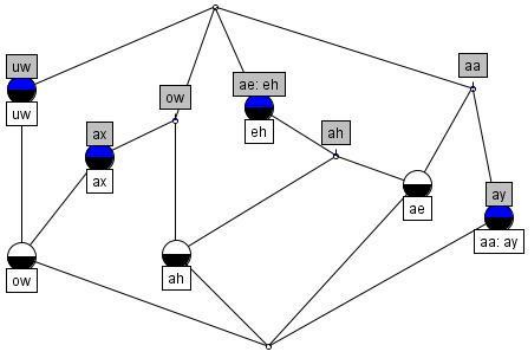

Figura 88: Retículo de confusión del subconjunto ISOLET VOCALES MFCC Noisy con φ= -0.062422 y 12 conceptos.

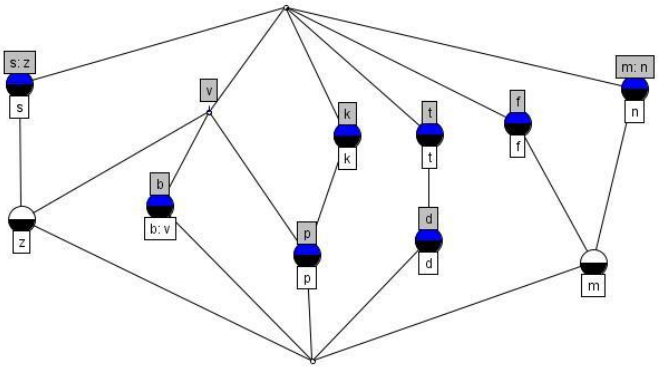

Figura 89: Retículo de confusión del subconjunto ISOLET-M&N55 MFCC Noisy y los experimentos de M&N con φ= 0.263326 y 13 conceptos.

8.7 FRECUENCIA 3000-5000Hz.

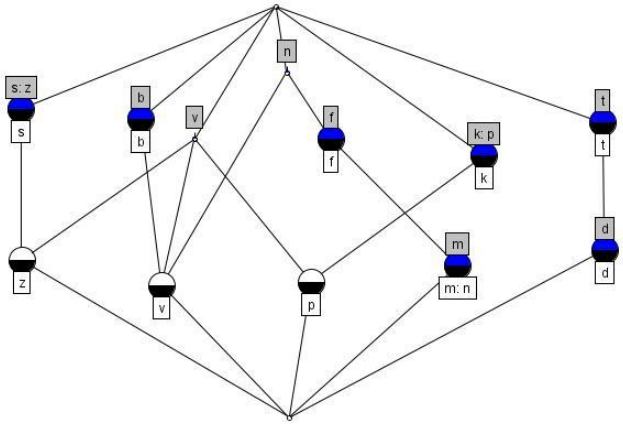

Figura 90: Retículo de confusión del subconjunto ISOLET CONSONANTES MFCC Noisy con phi -0,403932 y 19 conceptos.

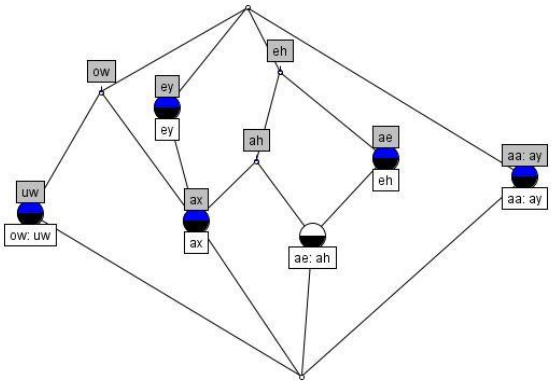

Figura 91: Retículo de confusión del subconjunto ISOLET VOCALES MFCC Noisy con φ= -0.191156 y 11 conceptos.

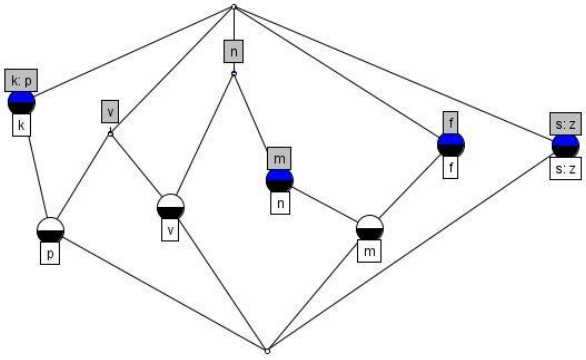

Figura 92: Retículo de confusión del subconjunto ISOLET-M&N55 MFCC Noisy con φ= -0.129450 y 11 conceptos.

8.8 CONCLUSIONES

A continuación comentamos en líneas generales el comportamiento con ruido proporcionando algunos ejemplos de los retículos en los que los efectos se observan más claramente.

En cuanto a los borrados de altas frecuencias tenemos que decir que las bandas de 200-300, 200-400 y 200-600 tienen los errores menos estructurados y difieren más de los que predice la teoría fonética. Ejemplos son los de las semi-vocales que se confunden habitualmente con las líquidas y las nasales (/w/ con /n/ -Figura 72-, /y/ con /l/ y /n/ -Figura 75-). Como estas consonantes no están en el conjunto analizado por M&N55 no podemos comprobar si este hecho se produce también en RHH.

En los borrados de bajas frecuencias constatamos que la separación entre los canales previstos por M&N55 ya no es tan nítida como en los casos sin ruido observando habitualmente nexos entre las fricativas y las oclusivas por una parte (por ejemplo, en la Figura 87, Figura 89, Figura 90 o Figura 92), las líquidas y las nasales (por ejemplo, en la Figura 87).

Al respecto de las vocales observamos cómo con los borrados de frecuencias altas aparecen las confusiones con /iy/ (que no aparece habitualmente por estar perfectamente clasificada y que además es la más frecuente en la base de datos –véanse los histogramas en la Figura 24 y Figura 30-) por una parte y por otra, con /ow/ y /uw/ que son, en gran medida las que acompañan a las semi-vocales en las pronunciaciones de las letras del alfabeto y que por tanto, creemos que están más relacionadas con la tarea particular que resolvemos que con la teoría fonética. Además, observamos en los histogramas que la presencia de estas vocales es menor que /ay/, /ey/ y /eh/ que son las mayoritarias junto con /iy/. Por eso, lo habitual es que estas tres vocales aparezcan con etiquetas grises en la parte alta de los retículos indicando que en muchas ocasiones se asignan estas etiquetas a las vocales /aa/, /ax/, /ae/ y /ah/ (con presencia minoritaria en la base de datos) en especial en presencia de ruido (por ejemplo, Figura 76, Figura 88 y Figura 91). La confusión de estas últimas vocales entre sí es constante en todos los casos analizados.