SISTEMA DE CONTROL DE ACCESO MEDIANTE EL RECONOCIMIENTO DE HABLA CONTINUA

GEYDY MOGOLLON PIÑA

ALEXANDER PALACIO UTRIA

DIRECTORA:

MARGARITA UPEGUI FERRER

UNIVERSIDAD TECNOLÓGICA DE BOLIVAR

MINOR EN AUTOMATIZACIÓN INDUSTRIAL

CARTAGENA

2003

LISTA DE ABREVIACIONES

ADC	Convertidor Analógico – Digital
NN	Red Neuronal
ASR	Reconocimiento automático de voz
DCT	Transformada del coseno discreta
DFT	Transformada Discreta de Fourier
DTW	Alineamiento Temporal
EM	*Expectation-Maximization*
FIR	Respuesta al Impulso Finito
HMM	Modelo Oculto de Markov
IDFT	Transformada Inversa Discreta de Fourier
LP	Predicción Lineal
LPC	Codificación por Predicción Lineal
MFCC	Coeficiente Cepstral Mel – Frecuency
MLP	Perceptron Multicapa
Ms-TDNN	Red Neuronal Time Delay Multiestados
TDNN	Red Neuronal Time Delay

TABLA DE CONTENIDO

LISTA DE FIGURAS

INTRODUCCION

El principal interés en el estudio de sistemas de reconocimiento de voz es proveer una forma más natural de interacción humano-computadora. Para este modelo de interacción se pretende la creación de interfaces centradas en las necesidades del usuario aprovechando una de las habilidades que tiene el humano para comunicarse: el habla.

El objetivo principal de estas investigaciones es hacer una descripción de los sistemas de reconocimiento del lenguaje hablado capaces de reconocer voz de manera espontánea, continua y sin restricciones.

El paso de un tipo de representación a otra no está exento de dificultades, dada la naturaleza variable de la onda sonora que producimos al hablar. Por una parte, una misma unidad lingüística por ejemplo un fonema no tiene idéntica manifestación sonora en todos los contextos en que puede aparecer y, a su vez, en un determinado segmento de la onda sonora se encuentra información acústica correspondiente a más de una unidad lingüística. Por otra parte, cada hablante posee características individuales algunas de ellas intrínsecas como el sexo, otras variables como la edad, el estado físico o el estado emocional y rasgos que señalan su pertenencia a un grupo social y a una zona geográfica, al tiempo que puede llegar a dominar una amplia gama de estilos o diferencias en el habla asociadas a una situación comunicativa particular. Mientras que el sistema humano de procesamiento del habla está preparado para responder a tales variaciones y también para realizarlas, la producción y el reconocimiento automáticos por parte de un ordenador suponen la capacidad de tratar adecuadamente los tipos de variación que se acaban de mencionar.

La necesidad de desarrollar aplicaciones informáticas que, de alguna manera, realicen las dos operaciones básicas de producir y reconocer, obliga a responder a una serie de preguntas que estimulan nuevas investigaciones en el campo de la fonética y a formalizar en sistemas de reglas explícitas los conocimientos ya adquiridos. A pesar de las dificultades anteriormente mencionadas, la posibilidad de comunicarse oralmente con los ordenadores ofrece una serie de ventajas: La libertad de movimientos y la facilidad para llevar a cabo otras tareas al no tener que depender de una pantalla y un teclado, la potencial simultaneidad del habla con otros sistemas de comunicación, la posibilidad casi universal de acceso remoto mediante el teléfono y, especialmente, el hecho de que el habla constituye nuestro sistema de comunicación más natural y, por lo tanto, su utilización contribuye a acercar el mundo de las tecnologías de la información a una amplia gama de personas, incluyendo aquellas que, por razones diversas como la edad, el nivel cultural o determinadas discapacidades, tienen dificultades con la lectura, la escritura o con el uso de teclados y monitores.

Al abordar el campo del reconocimiento del habla (*Automatic Speech Recognition, ASR)*, es necesario realizar una primera diferenciación entre 'reconocimiento' y 'comprensión'. En el campo de las tecnologías del habla, suele trabajarse en el nivel del reconocimiento, es decir, en la conversión de una señal acústica -el habla- en una cadena de símbolos o de caracteres correspondientes a esta señal. Así puede generarse un texto escrito o controlar el funcionamiento de un sistema informático, utilizarse como interfaz de control de un robot, para incorporar nuevas reglas a un sistema experto, etc. pero no se consigue la comprensión, si por ello se entiende la extracción de información semántica.

1. PRODUCCION Y PERCEPCION DE LA VOZ

El uso que los seres humanos hacen del lenguaje y los mecanismos envueltos en el proceso por el cual las personas pueden transmitir sus ideas o conceptos a otras son temas relacionados con la comunicación del lenguaje humano. No es necesario recalcar que estos mecanismos son muy importantes para el desarrollo de las tecnologías relacionadas con el lenguaje humano.

La mayoría de estos procesos son cubiertos por la lingüística. No en vano la lingüística es el estudio del lenguaje humano (su estructura, su significado, su utilización para comunicar ideas, como esta formado y como es decodificado). Sin embargo las disciplinas más importantes relacionadas con el reconocimiento de habla continua (CSR de sus siglas en inglés) son la fonética y la fonología en la producción del habla y la fisiología y Psicología en la percepción del habla. Tanto la producción como la percepción del lenguaje humano son extremadamente importantes para tratar de entender la relación entre la forma de onda que el receptor escucha y la posterior decodificación del mensaje. Finalmente el código utilizado por el locutor y el receptor debe ser un lenguaje el cual puede ser descrito por un número finito de sonidos distinguibles.

1.1 PRODUCCION DE LA VOZ

El habla es el producto acústico de determinados movimientos voluntarios de los aparatos respiratorios y masticatorio. Esta capacidad se tiene que aprender normalmente en el primer año de nuestras vidas y es desarrollada, controlada y

mantenida por la retroalimentación acústica del mecanismo auditivo**.** Toda esta información es recibida y procesada por el sistema nervioso central.

Como cualquier otro sonido el habla es el resultado de una variación de presión del aire. Esta variación es causada por la constricción que el aire encuentra cuando fluye de los pulmones a través del tracto vocal hacia el exterior. Los pulmones actúan de tal forma que producen un flujo de aire a través de la traquea, laringe (donde se encuentran ubicadas las cuerdas vocales) ya sea al tracto vocal y expulsado por los labios o a la cavidad nasal y expulsados por las fosas nasales dependiendo de la abertura del velo del paladar.

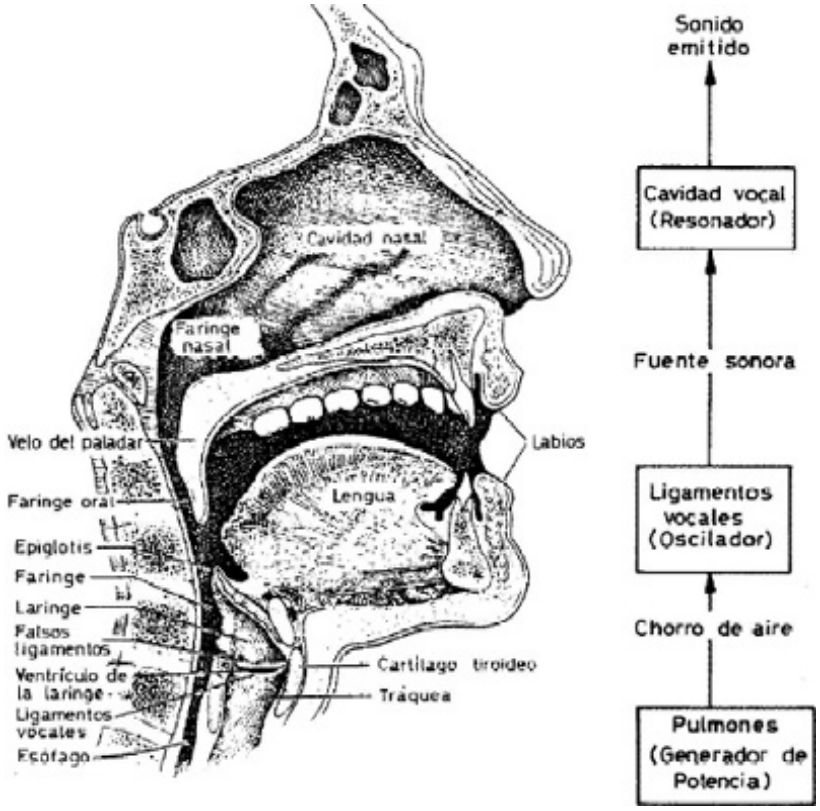


Figura 1.1 Fisiología del aparato fonador humano

El tracto vocal cuyos extremos son las cuerdas vocales y los labios pueden ser considerados como un tubo acústico con una sección de área no uniforme ya que esta área depende de la posición de los articuladores (labios, mandíbula,

lengua, etc.) para diferentes configuraciones de estos articuladores se produce un sonido diferente.

Pero no solo la configuración de los articuladores producen la diferencia entre sonidos ya que también es importante la naturaleza de la fuente de excitación.

Por un lado, la fonación es el proceso por el cual una excitación cuasi-periódica es producida en la glotis.

El volumen de aire por unidad de tiempo es mas o menos proporcional al área de la glotis y en un tono normal la forma de esta función es aproximadamente un pulso periódico triangular con un factor de utilidad entre 0.3 y 0.7. Ya que el espectro presenta muchos armónicos y debido a la señal triangular, la amplitud de ellos decrece en doce decibeles por octava. Sin embargo, debido a la pequeña abertura de las cuerdas vocales la impedancia acústica de la fuente glotal es alta comparada con la del tracto vocal. Por lo tanto las variaciones en el tracto vocal tienen poca influencia sobre la fuente.

En términos eléctricos pueden ser considerados como un generador de corriente constante en un circuito que es variable en el tiempo.

Por otro lado, otra excitación puede ser producida por la turbulencia del aire creado por los articuladores en alguna parte del tracto vocal. Por lo cual se crea un fluido acústico que proporciona una excitación incoherente con la fuente. Esta fuente puede ser considerada como primera aproximación independiente de la anterior y es la fuente que hace los continuos sonidos sordos. Sin embargo, medidas indirectas y algunas teorías sugieren que el espectro de estos sonidos tiene una distribución uniforme y que las cavidades resonantes posteriores a la constricción son usualmente más influyentes que las anteriores en el espectro de onda.

A parte de las clases de fuente utilizadas en la producción del habla la forma del espectro resultante es un producto de la resonancia (llamadas *formantes*) que son producidos en las cavidades resonantes y el efecto de radiación. Pero el punto es que la influencia de estos resonadores depende de la configuración de los articuladores que cambian el área de sección transversal o en el caso de la cavidad nasal el velo del paladar hace que fluya el aire o no a través de este.

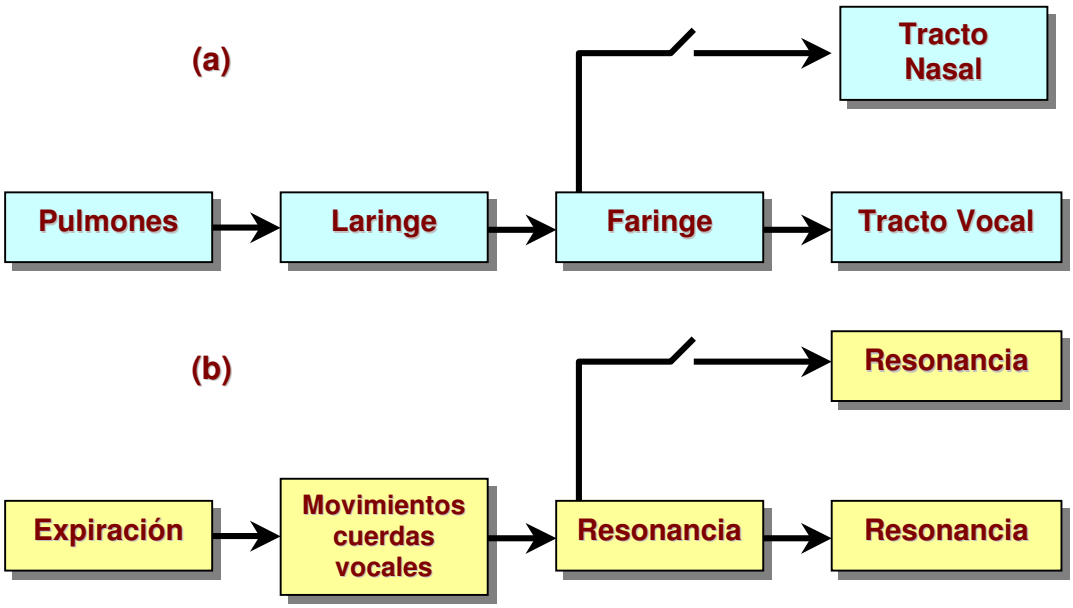


Figura 1.2. Diagrama Funcional del Aparato Fonador

1.1.1 Clasificación de los sonidos de la lengua

Los sonidos de la lengua se clasifican en:

1.1.1.1 Sonidos Sordos. Se realizan sin la intervención de las cuerdas vocales (posición de respiración),donde el flujo de aire procedente de los pulmones se vuelve más rápido produciendo fricciones y turbulencias las cuales se traducen en vibraciones en aquellos puntos en que por ocasionarse un estrechamiento la velocidad del fluido alcanza máximos. Si el punto de estrechamiento o

articulación se encuentra próximo al exterior (labios y dientes) el resto de la estructura influye escasamente en el sonido emitido, en caso contrario (zona velar o medial) la posición de los restantes órganos puede modular bastante el resultado. La forma de onda de este tipo de sonidos presenta un escaso carácter repetitivo recordando el ruido aleatorio.

Los sonidos sordos son típicamente consonánticos aunque también hay consonantes sonoras.

1.1.1.2 Sonidos Sonoros**.** Estos se realizan con la intervención de las cuerdas vocales, que obstruyen el paso del flujo de aire durante breves instantes para ceder expeliendo breves pulsos glotales. Este tren de pulsos es modificado por los órganos de articulación que vienen posteriormente añadiendo una cierta codificación. Esta codificación consiste en el resalte de determinadas frecuencias contenidas en el pulso glotal original o en su eliminación. El sonido así emitido tiene una forma relativamente periódica. Un ejemplo de sonidos sonoros lo constituyen las vocales.

1.1.2 Punto de Articulación

Son los puntos de estrechamiento en las cavidades supraglóticas, antes mencionados, para los sonidos sordos y sonoros. Estos son controlados por los órganos móviles contra los órganos inmóviles de su misma zona.

1.1.3 Modo de Articulación

Define la forma concreta en que la articulación puede tener lugar. Se puede dar de dos posibilidades: modo espirado y modo no espirado. El modo espirado se produce bajo un flujo de aire de origen pulmonar y es el modo habitual en la mayoría de las lenguas. El modo no espirado se produce por compresión o enrarecimiento del aire en la cavidad bucal, que sufre dos cierres, uno de los cuales cede bruscamente con entrada o salida de flujo aéreo.

Dentro del modo espirado y dependiendo del tipo de aproximación practicada por los órganos articulatorios podemos tener varias submodalidades: consonantes, sonantes y vocales.

1.1.3.1 Consonantes. Corresponden al máximo cierre del tracto vocal en algún punto concreto. Pueden ser: Oclusivas y fricativas.

1.1.3.1.1 Oclusivas. Se produce una obstrucción total al paso del flujo aéreo, pero solamente durante breves instantes, dando lugar a una explosión posterior a la apertura del tracto vocal. Su duración por lo general es de entre 10 a 30 ms.

1.1.3.1.2 Fricativas. Se produce una obstrucción incompleta del flujo, que busca los pequeños resquicios residuales que puedan existir, aumentando enormemente la velocidad del fluido en la zona. Pueden tener una duración superior a los 100ms.

1.1.3.2 Sonantes. Son aquellas articulaciones en las cuales no se produce estrechamiento suficiente para que aparezca fricción, y sólo se realiza una modificación por cierre parcial o derivación. Estas pueden ser: nasales, líquidas, vibrantes y deslizantes.

1.1.3.2.1 Nasales. La cavidad nasal se abre al paso del flujo aéreo, produciendo una oclusión total o casi total en algún punto de articulación de la cavidad oral. La radiación es bastante menor en amplitud que la que se produciría por el tracto vocal.

1.1.3.2.2 Líquidas. Se basan en una obstrucción incompleta alrededor del tercio anterior de la lengua. El ápice bloquea total o parcialmente al flujo glotal, pero los bordes laterales de la lengua permiten en el paso por el espacio que queda a ambos lados de la lengua.

1.1.3.2.3 Vibrantes. Variante de las líquidas, en las que el ápice realiza bruscas aperturas y cierres de su contacto con la zona alveolar o prepalatal, junto con la oclusión total o parcial del flujo lateral por medio de las zonas laterales de la lengua.

1.1.3.2.4 Deslizantes. Son sonidos fuertemente relacionados con la evolución de un sonido a otro en la articulación de un diptongo. En tal caso, uno de los sonidos de tipo vocálico se consonantiza por cierre excesivo. Para su articulación requieren siempre de la presencia de una vocal.

1.1.3.3 Vocales. Corresponden la mayor apertura del tracto vocal. La energía emitida correspondiente a una vocal suele ser mucho mayor que la correspondiente a consonantes o sonantes. En ningún caso llegan a presentar una obstrucción o estrechamiento superior a un 70% de su sección media en ningún punto. La apertura en las vocales permite clasificarlas en cerradas, semicerradas, semiabiertas y abiertas. Otro rasgo de las vocales que permite clasificarlas es su punto de articulación o de máximo estrechamiento: anteriores, medias y posteriores.

1.2 PERCEPCION DE LA SEÑAL DE VOZ

En esta sección del capitulo se explica cómo se percibe un sonido desde el exterior hasta llegar al cerebro humano, donde cada una de las partes del oído son esenciales para la percepción de este.

1.2.1 Estructura del oído.

El oído se encuentra dividido en tres zonas: externa, media e interna. La mayor parte del oído interno está rodeada por el hueso temporal.

El oído externo es la parte del aparato auditivo que se encuentra en posición lateral al tímpano o membrana timpánica. Comprende la oreja o pabellón auricular o auditivo (lóbulo externo del oído) y el conducto auditivo externo, que mide tres centímetros de longitud aproximadamente.

El oído medio se encuentra situado en la cavidad timpánica llamada caja del tímpano, cuya cara externa está formada por la membrana timpánica, o tímpano, que lo separa del oído externo. Incluye el mecanismo responsable de la conducción de las ondas sonoras hacia el oído interno. Es un conducto estrecho, o fisura, que se extiende unos 15 mm en un recorrido vertical y otros 15 mm en recorrido horizontal aproximadamente. El oído medio está en comunicación directa con la nariz y la garganta a través de la trompa de Eustaquio, que permite la entrada y la salida de aire del oído medio para equilibrar las diferencias de presión entre éste y el exterior. Hay una cadena formada por tres huesos pequeños y móviles que atraviesa el oído medio. Estos tres huesos reciben los nombres de martillo, yunque y estribo. Los tres conectan acústicamente el tímpano con el oído interno, que contiene un líquido.

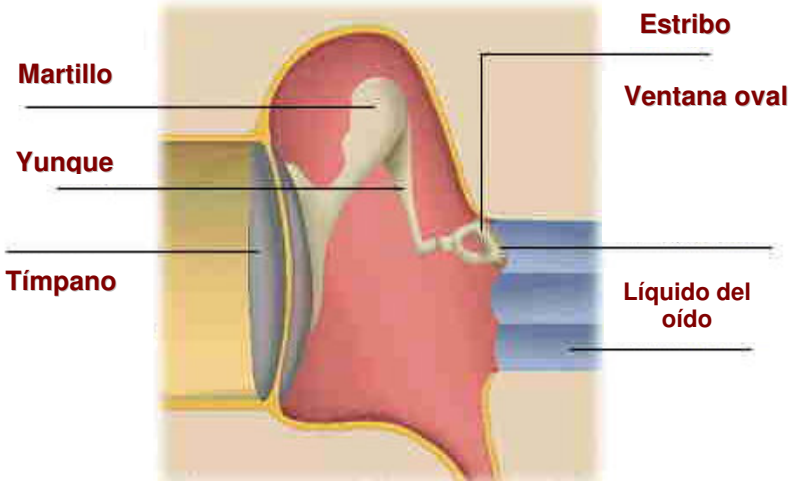


Figura 1.3. Estructura del Oído

El oído interno, se encuentra en el interior del hueso temporal que contiene los órganos auditivos y del equilibrio, que están inervados por los filamentos del nervio auditivo.

Está separado del oído medio por la *fenestra ovalis*, o ventana oval. El oído interno consta de una serie de canales membranosos alojados en una parte densa del hueso temporal, y está dividido en: cóclea (en griego, 'caracol óseo'), vestíbulo y tres canales semicirculares. Estos tres canales se comunican entre sí y contienen un fluido gelatinoso denominado endolinfa.

1.2.2 Percepción de las ondas sonoras.

Para poder percibir las ondas sonoras, el cuerpo cuenta con un complejo mecanismo formado por el oído externo, el oído medio y el interno.

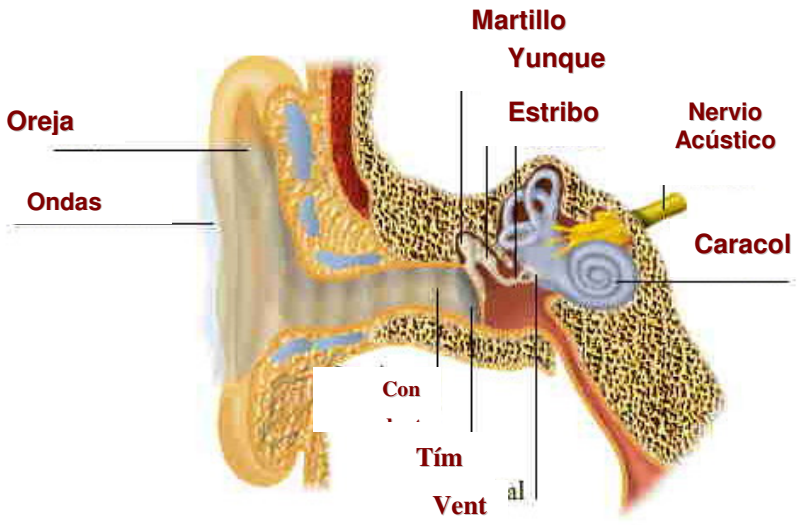


Figura 1.4. Como viajan las ondas sonoras a través del oído

La aurícula recoge las ondas sonoras y las conduce por el conducto auditivo. Las ondas sonoras chocan contra el tímpano, que, como consecuencia, vibra. Las vibraciones se transmiten gracias a una cadena de tres huesecillos: El

martillo, el yunque y el estribo. Las vibraciones pasan por la ventana oval y llegan al caracol, ya en el interior del oído interno. Allí las vibraciones se convierten en impulsos nerviosos, estos transcurren por el nervio acústico hasta el cerebro, donde son interpretados como sonidos.

Las ondas sonoras son producidas por cambios en la presión del aire, estas son transmitidas a través del canal auditivo externo hacia el tímpano, en el cual se produce una vibración. Estas vibraciones se comunican al oído medio mediante la cadena de huesecillos (martillo, yunque y estribo) y, a través de la ventana oval, hasta el líquido del oído interno. El movimiento de la endolinfa que se produce al vibrar la cóclea, estimula el movimiento de un grupo de proyecciones finas, similares a cabellos, denominadas células pilosas. El conjunto de células pilosas constituye el órgano de Corti. Las células pilosas transmiten señales directamente al nervio auditivo, el cual lleva la información al cerebro. El patrón de respuesta de las células pilosas a las vibraciones de la cóclea codifica la información sobre el sonido para que pueda ser interpretada por los centros auditivos del cerebro.

El rango máximo de audición en el hombre incluye frecuencias de sonido desde 16 hasta 28.000 ciclos por segundo. El menor cambio de tono que puede ser captado por el oído varía en función del tono y del volumen. Los oídos humanos más sensibles son capaces de detectar cambios en la frecuencia de vibración (tono) que correspondan al 0,03% de la frecuencia original, en el rango comprendido entre 500 y 8.000 vibraciones por segundo. El oído es menos sensible a los cambios de frecuencia si se trata de sonidos de frecuencia o de intensidad bajas.

La sensibilidad del oído a la intensidad del sonido (volumen) también varía con la frecuencia. La sensibilidad a los cambios de volumen es mayor entre los 1.000 y los 3.000 ciclos, de manera que se pueden detectar cambios de un decibelio. Esta sensibilidad es menor cuando se reducen los niveles de intensidad de sonido.

Las diferencias en la sensibilidad del oído a los sonidos fuertes causan varios fenómenos importantes. Los tonos muy altos producen tonos diferentes en el oído, que no están presentes en el tono original. Es probable que estos tonos subjetivos estén producidos por imperfecciones en la función natural del oído medio. Las discordancias de la tonalidad que producen los incrementos grandes de la intensidad de sonido, es consecuencia de los tonos subjetivos que se producen en el oído. Esto ocurre, por ejemplo, cuando el control del volumen de un aparato de radio está ajustado. La intensidad de un tono puro también afecta a su entonación. Los tonos altos pueden incrementar hasta una nota de la escala musical; los tonos bajos tienden a hacerse cada vez más bajos a medida que aumenta la intensidad del sonido. Este efecto sólo se percibe en tonos puros. Puesto que la mayoría de los tonos musicales son complejos, por lo general, la audición no se ve afectada por este fenómeno de un modo apreciable. Cuando se enmascaran sonidos, la producción de armonías de tonos más bajos en el oído puede amortiguar la percepción de los tonos más altos. El enmascaramiento es lo que hace necesario elevar la propia voz para poder ser oído en lugares ruidosos.

2. SISTEMAS DE CONTROL DE ACCESO

Entre los sistemas de control de acceso basados en características biométricas se pueden mencionar, entre otros, aquellos basados en la voz, la retina y las huellas dactilares. En ciertas circunstancias estos sistemas son considerados más seguros y personales que los convencionales (tarjetas magnéticas, *password*, etc.).

Un sistema de control de acceso basado en el reconocimiento de voz se divide en dos áreas: La Identificación de Locutor, y la Verificación de Locutor (VL). Un sistema de identificación de locutor asignará al usuario en cuestión la identidad del individuo registrado que mejor se aproxime a las características de la señal de voz. Por otra parte, un sistema de VL deberá decidir si la persona que declara una cierta identidad es o no quien dice ser (Doddington, 1985; Furui, 1994).

La pronunciación emitida por un locutor cualquiera es comparada con el modelo del cliente cuya identidad fue declarada. De esta forma, si el modelo de locutor y la pronunciación coinciden dentro de los límites permitidos (umbral de decisión), la identidad será aceptada y en caso contrario será rechazada. Como se puede ver, en este tipo de sistemas sólo existen dos respuestas posibles: Aceptar o rechazar al locutor. Esta lleva a cuatro casos posibles, dos correctos y dos errados:

- Aceptar un locutor registrado.
- Rechazar un impostor.

- Aceptar un impostor.
- Rechazar un locutor registrado.

Los dos primeros casos corresponden a respuestas correctas por parte del sistema de VL, mientras que las dos últimas opciones son erradas. Estos errores corresponden, respectivamente, a los tradicionalmente denominados error de falsa aceptación y error de falso rechazo. La tasa donde estos errores se igualan (EER-*Equal Error Rate*) es comúnmente utilizada como medida de desempeño en VL y otros sistemas biométricos.

Todos los sistemas de VL cuentan con una base de datos de usuarios registrados, denominados clientes. Esta base de datos está compuesta por modelos que representan las características del habla de cada uno de los clientes. Estos modelos se consiguen mediante el procesamiento de sesiones de entrenamiento en las cuales el usuario del sistema pronunciará varias frases.

Existen diversos tipos de sistemas de VL, entre ellos se pueden distinguir los sistemas de texto dependiente y los de texto independiente. Los sistemas de texto dependiente requieren que el usuario pronuncie una palabra o frase determinada por el sistema. Los sistemas de texto independiente están preparados para realizar el proceso de VL cualquiera sea la palabra o frase pronunciada. Se pueden distinguir dentro de cada uno de estos tipos de sistema aquellos de pronunciación continua o los de palabra aislada. En estos últimos las palabras deberán estar separadas entre sí por pequeños instantes de silencio; nuestro campo de investigación se basa en los sistemas de pronunciación continua.

2.1 ETAPAS EN UN SISTEMA DE VL

Las etapas que contempla un sistema de VL, tanto en su fase de entrenamiento como de verificación, se pueden ver en la Figura 2.

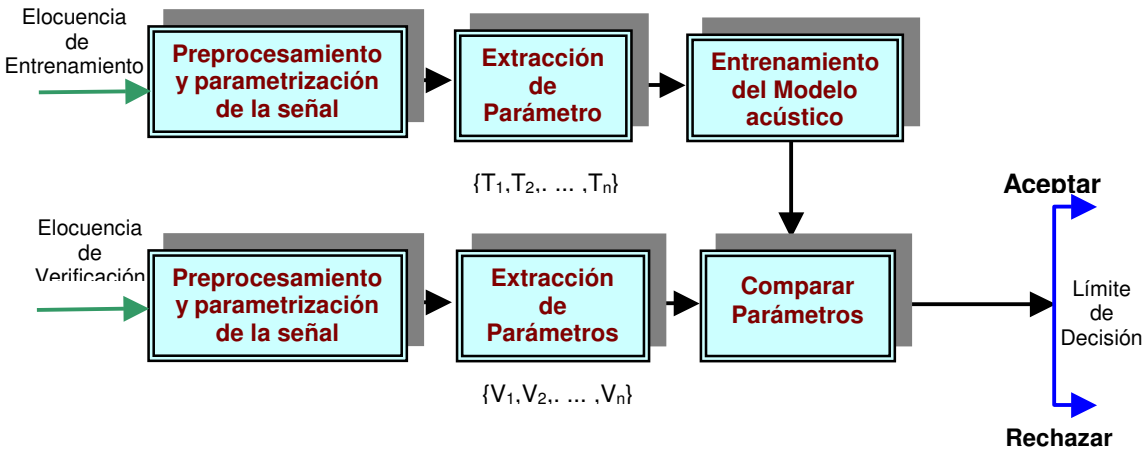


Figura 2.1. Diagrama de bloques de un sistema de VL

La señal de voz correspondiente a las sesiones de entrenamiento es sometida a una etapa de pre-procesamiento que consiste en un filtrado anti-aliasing, conversión análoga-digital, detección de inicio-y-fin de la señal de voz útil, y pre-énfasis. Posteriormente, en la etapa de extracción de parámetros, la señal se divide en cuadros o *frames*, se realiza una ponderación al interior de éstos por una ventana (tipo Hamming) para luego realizar un análisis espectral que permitirá obtener los parámetros deseados. Con estos parámetros el sistema de VL deberá ser capaz de entrenar un modelo que represente al cliente.

Por su parte, las elocuciones de verificación serán sometidas al mismo proceso, salvo que los parámetros obtenidos serán comparados con los modelos del cliente para tomar la decisión adecuada.

2.2 FUNDAMENTOS DEL RECONOCIMIENTO DE HABLA CONTINUA

El proceso de reconocimiento automático de voz consiste en: primero obtener y digitalizar la señal de voz; segundo, extraer un conjunto de características esenciales de la señal para después introducirlas a un clasificador el cual se encarga de obtener probabilidades para cada conjunto de características que se introduzcan[1].

Una vez obtenidas estas probabilidades y con la ayuda de una estructura que nos dé las pronunciaciones posibles que deseamos reconocer, se debe realizar un algoritmo de búsqueda para encontrar la secuencia permitida más probable, y así finalizar el reconocimiento encontrando las palabras que se deseaban reconocer. En la figura 0.1 se muestra la estructura general de un reconocedor automático de voz.

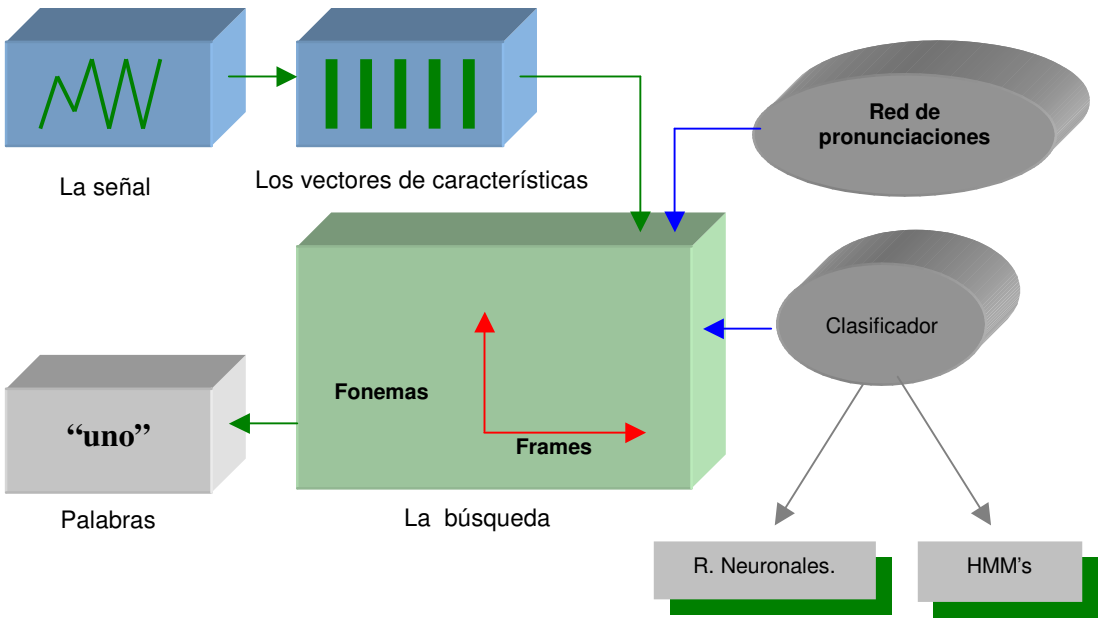


Figura 2.2. Esquema de un sistema general de reconocimiento de voz

A continuación se da una breve descripción de cada una de las fases necesarias para llevar a cabo el reconocimiento de voz:

[1] PELTON, Gordon E. Voice Processing. San Francisco: McGraw-Hill,1993. p 83

- **La señal de Voz**. La señal de voz se muestrea generalmente a una frecuencia de 16 KHz para señales provenientes de micrófonos o 8 KHz para señales telefónicas. Esto proporciona una secuencia de valores de amplitud en el tiempo.

- **Análisis de la señal de voz**. La señal de voz debe transformarse y comprimirse inicialmente. Existen muchas técnicas que nos permiten extraer las características de la señal y comprimir los datos obtenidos en un factor de 10 si perdida de información relevante.

- **Tramas de voz (*Frames*)**. El resultado del análisis de la señal de voz es una secuencia de tramas de la señal de voz, típicamente a un intervalo de 10ms, con cerca de 16 coeficientes por trama. Estas tramas pueden ser aumentadas por su primera y/o segunda derivada, proporcionando información explicita sobre la dinámica de la señal de voz; Esto típicamente permite un mejoramiento en el proceso.

- **Modelos Acústicos.** Con el fin de analizar el contenido acústico de las tramas de la señal, se hace necesaria la implementación de un modelo acústico. Existe gran cantidad de modelos acústicos, variando en su representación, contexto dependiente entre otras propiedades.

  El más simple se basa en plantillas, el cual se almacenan muestras de la unidad del habla que se desea modelar, por ejemplo la grabación de una palabra. Una palabra desconocida se puede reconocer simplemente comparándola con una plantilla conocida previamente, encontrando la relación mas adecuada.

  Una representación más flexible se utiliza para sistemas más robustos, basado en modelos acústicos entrenados, o estados. En este caso, cada palabra es modelada por una secuencia de estados entrenables, y cada estado indica el sonido que parece escucharse en ese segmento de la palabra, utilizando una distribución probabilística sobre el espacio acústico.

La distribución de probabilidad puede modelarse paramétricamente, asumiendo que tienen una forma simple (como por ejemplo una distribución Gaussiana) y luego tratando de encontrar los parámetros que la describen, esta etapa también puede modelarse de forma no paramétrica, representando la distribución directamente (con un histograma sobre una cuantización del espacio acústico o como se describirá en capítulos posteriores mediante una red neuronal).

Durante el entrenamiento, los modelos acústicos son modificados incrementalmente con el fin de optimizar su comportamiento. Durante la fase de prueba los modelos acústicos son muy poco cambiados.

- **Análisis acústico**. El análisis acústico es implementado utilizando un modelo acústico sobre cada trama de la señal de voz, proporcionando una matriz de *scores*, como se muestra en la figura. Estos *scores* son calculados de acuerdo al tipo de modelo acústico que esta siendo utilizado. Para los modelos acústicos basados en plantillas, u score es típicamente la distancia Euclidiana entre la trama de la plantilla y la trama desconocida. Para modelos acústicos basados en estados, un *score* representa una probabilidad de emisión, ejemplo la probabilidad del estado actual de generar la trama actual, determinado por la función paramétrica o no paramétrica de estado.

- **Alineamiento temporal.** Las tramas *scores* son convertidas a una secuencia de palabras, mediante la identificación una secuencia de modelos acústicos, representando una secuencia de palabras validas, las cuales proporcionarán un camino de alineamiento, a través de una matriz, como la que se ilustra en la figura. El proceso de búsqueda del mejor camino es llamado Alineamiento temporal.

  El alineamiento puede ser mejorado eficientemente por métodos de programación dinámica, un algoritmo general el cual utiliza solo las

limitaciones de caminos locales, y el cual cumple con los requerimientos de espacio y tiempo. Este algoritmo general tiene dos variantes principales la distorsión dinámica temporal (DTW) y la búsqueda de Viterbi que son se expondrán en el capítulo de técnicas clásicas de reconocimiento de voz.

En un sistema basado en estados, el camino de alineamiento optimo conduce a una segmentación en la secuencia de la palabra, como tal indica cuales tramas son asociadas con cada estado. Esta segmentación puede utilizarse para generar etiquetas para el entrenamiento recursivo del modelo acústico sobre las correspondientes tramas.

3. PREPROCESAMIENTO DE LA SEÑAL DE VOZ

Para poder realizar un reconocimiento del habla, el primer paso a seguir es la adquisición de los datos necesarios para trabajar, es decir, capturar el sonido y cuantificarlo para que pueda ser procesado mediante una computadora. Para la función de capturar el sonido se pueden utilizar distintos tipos de micrófonos los cuales, dependiendo de su calidad, pueden llegar a tener una sensibilidad mayor que la del oído humano. Éstos impulsos, de naturaleza analógica, han de ser transformados mediante un conversor analógico-digital que se encargará de realizar el muestreo y la cuantificación de la señal para luego proceder con la extracción de características de esta señal de voz.

3.1 ADQUISICIÓN DE LA SEÑAL DE VOZ

En el proceso de adquisición de la señal de voz, se utiliza un transductor (normalmente un micrófono) capaz de convertir la señal de presión de aire en una señal eléctrica. Entonces, esta señal analógica después de ser filtrada, es muestreada a una frecuencia de muestreo que obedezca al criterio de Nyquist: $Bw \leq \frac{1}{2} F_s$, donde Bw es el ancho de banda de la señal y F_s la frecuencia de muestreo. Además los valores de las muestras deben ser cuantificados para que estos puedan ser almacenados en dispositivos digitales.

Los dispositivos que permiten esta función son los llamados Convertidores Analogico/Digital (A/D o ADC), su objetivo es proporcionar un dato muestreado con una relación Señal a Ruido (SNR) tan alta como sea posible, porque, debido a esta conversión, la señal original se distorsiona ligeramente y esto tiene como consecuencia la suma de ruido a la señal.

Además en muchos reconocedores de habla, las frecuencias mas altas son amplificadas por filtros pre-énfasis, con el fin de resaltar las características acústicas de la señal en esta región, La etapa de preénfasis se encarga de suavizar el espectro y reducir las inestabilidades de cálculo asociadas con las operaciones aritméticas de precisión finita. Además se usa para compensar la caída de -6 dB que experimenta la señal al pasar a través del tracto vocal; un filtro digital de primer orden tipo FIR (*Finite Impulse Response*),cuya función de transferencia es:

$$H_{pre}(Z) = 1 - a_{pre} Z^{-1}$$

Donde a_{pre} esta en un rango típico de [0.4,1.0].

La representación mas adecuada para el análisis de la señal de voz son los espectrogramas.

Un espectrograma es una representación de la señal de voz de acuerdo a las variaciones de la energía, con respecto al tiempo y la frecuencia. El espectrograma contiene mucha información y releva las características acústicas especificas del habla.

Las bandas oscuras observadas en un espectrograma corresponden a las concentraciones de energía y llamadas formantes. Los formantes son las frecuencias en las que ocurre la resonancia de las vibraciones vocales.

Los espectrogramas son útiles para un análisis visual de la señal. sin embargo un reconocedor debe extraer de la señal acústica solo la información que requiere para poder reconocer una frase. Para ello la señal se muestrea a cierta frecuencia como se vio anteriormente, se cuantiza y posteriormente se crean vectores de características.

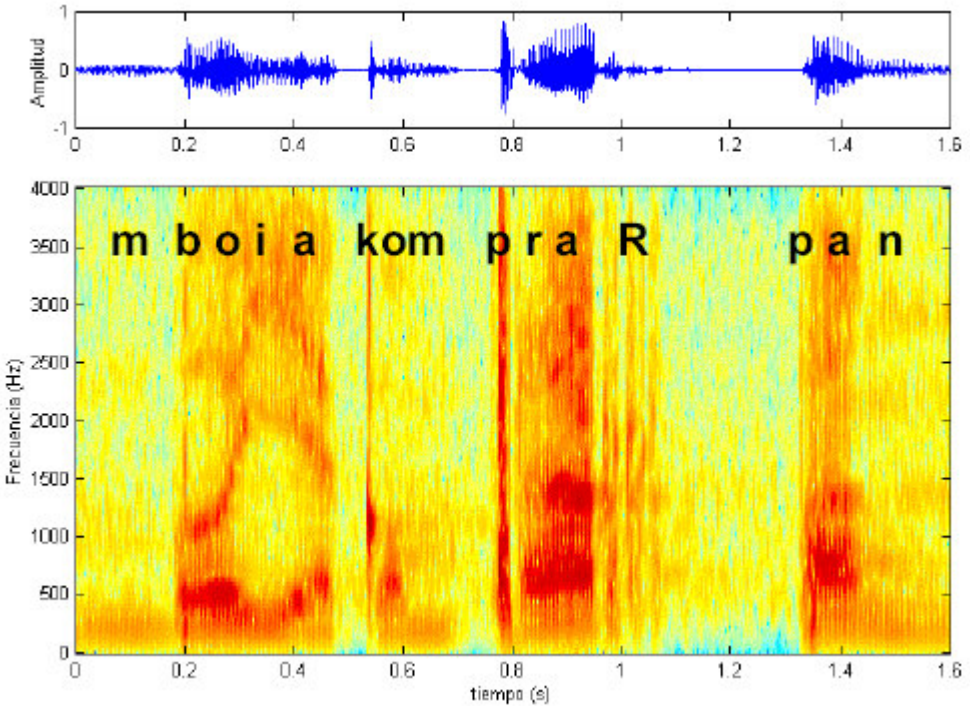


Figura 3.1. Espectrograma de la frase Voy a comprar Pan

3.2 ALGORITMOS DE EXTRACCIÓN DE CARACTERÍSTICAS

Como se comento en el primer capitulo de este trabajo, el oído interno tiene una zona de respuesta diferente para cada frecuencia. La forma del espectro S(f) de una señal es el resultado de la forma de la fuente de excitación G(*f*) y de las modificaciones sufridas en el tracto vocal V(f), donde S(*f*) = G(*f*)V(*f*).

Debido a esto, se han hecho muchos esfuerzos tratando de estimar la función de transferencia del tracto vocal, independientemente de las características de la fuente de excitación, ya que la envolvente del espectro, da una aproximación de la configuración del tracto vocal, por ende, la naturaleza del sonido de la voz.

Existen varias técnicas que permiten la extracción de las características de la señal de voz las cuales se describen a continuación.

3.2.1 Análisis Por Banco de Filtros

El uso de bancos de filtros digitales implementados inicialmente como filtros analógicos, ha sido históricamente la primera aproximación al procesamiento del habla.

La justificación de este análisis esta relacionada con funcionamiento del oído humano, especialmente con el órgano de Corti el cual presenta zonas de respuesta distintas a distintos tonos o frecuencias, debidas a la forma de este. Así, la respuesta del oído interno puede ser modelada por un banco de filtros.

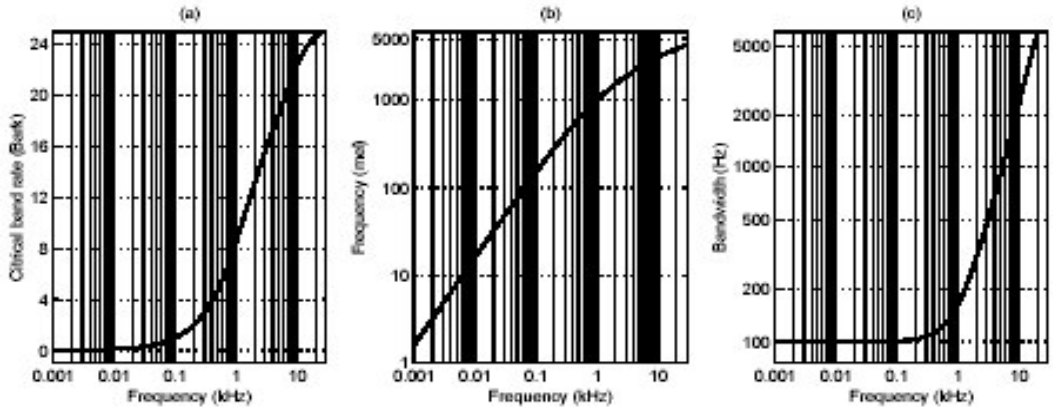


Figura 3.2. (a) y (b) Son diferentes escalas perceptuales, (b) Escala Mel, que es una aproximación de la escala de Bark (a), (c) es el ancho de banda critico que depende de la frecuencia central de filtro

La señal inicial se descompone en un conjunto discreto de muestras espectrales, que contienen una información similar a la que se presenta en los niveles superiores del sistema auditivo. Con objeto de aproximarse a la sensibilidad del oído humano, que no tiene una respuesta lineal en frecuencia existen diferentes escalas.

Estas escalas son más logarítmicas, donde se describen escalas perceptuales para el eje de frecuencia con el fin de simular el comportamiento del oído

humano: Esta la escala de Bark, aproximada por la escala de Mel (que es la más utilizada) y el ancho de banda del filtro $Bw_{critico}$.

$$Escala\ Bark = 13\arctan\left(\frac{0.76}{1000}\right) + 3.5\arctan\left(\frac{f^2}{(7500)^2}\right)$$

$$Escala\ Mel = 2595\log_{10}\left(\frac{1+f}{700}\right)$$

$$Bw_{critico} = 25 + 75\left[1 + 1.4\left(\frac{f}{1000}\right)^2\right]^{0.69}$$

La escala de Mel es la más utilizada en aplicaciones de tratamiento de la voz. Un banco de filtros está constituido por un conjunto de filtros cada uno de los cuales retiene la información de una serie determinada de frecuencias del espectro. A su vez cada filtro puede ponderar de manera diferente las frecuencias que quedan bajo su ámbito.

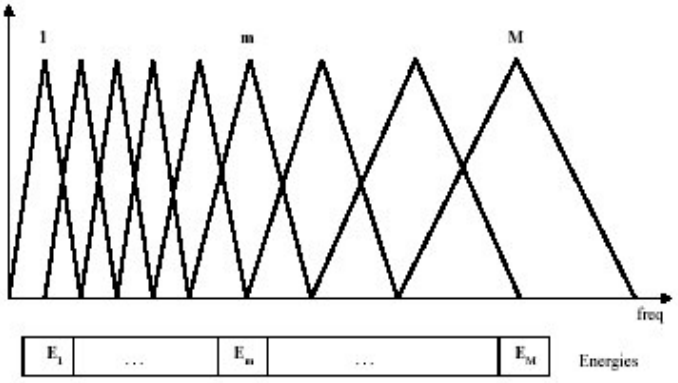


Figura 3.3. Diagrama de un banco de filtros a escala

Este tipo de técnica, generalmente se emplea de manera conjunta con otros métodos como son el cálculo de Coeficientes Cepstrales.

3.2.2 Transformada Discreta de Fourier

La transformada discreta de Fourier o DFT *(Discrete Fourier Transform)*, se define como:

$$X(k) = \sum_{k=0}^{N-1} x(n) e^{-j\left(\frac{2\pi kn}{N}\right)} \qquad k = 0,1,2,3,K,N-1$$

Donde N es el número de muestras de la ventana que se va a analizar.

Por su parte la DFT inversa o IDFT se define como:

$$X(k) = \sum_{k=0}^{N-1} x(n) e^{-j\left(\frac{2\pi kn}{N}\right)} \qquad k = 0,1,2,3,K,N-1$$

La motivación del uso de la DFT parte del hecho de la utilidad que tiene descomponer la señal de voz de partida en sus componentes en frecuencia.

Un aspecto importante si se quiere usar la DFT con señales de voz es que se debe asumir que al menos en periodos cortos de tiempo se cumple que la señal es estacionaria. En la realidad esto no es estrictamente así aunque podemos suponerlo[2].

La solución consiste en multiplicar la señal por una función ventana que sea 0 fuera de un determinado rango y reproducir el resultado de forma que tengamos un número de bloques iguales.

La ventana rectangular se define como: $w_n = \begin{cases} 1 & 0 \le n \le N \\ 0 & resto \end{cases}$

[2] OPPENHEIM, Alan V. Señales y Sistemas. 2ª Edición. México: Prentice Hall, 1997. p. 361 - 362

Sin embargo la utilización de esta ventana trae consigo, que en los puntos de inicio y fin exista una fuerte discontinuidad.

Para reducir el efecto de discontinuidad al mínimo debemos emplear tipos de ventana que tiendan a reducir a (0) los valores de las muestras en los extremos.

Aunque existe un buen número de tipos de ventana, la más común en el análisis de la voz es la que se conoce como ventana de Hamming (Figura 6):

$$w_n = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) & 0 \le n \le N \\ 0 & resto \end{cases}$$

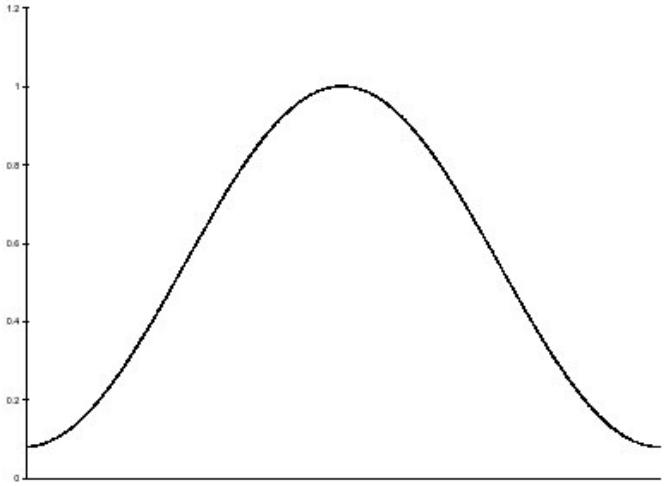


Figura 3.4. Ventana de Hamming

La complejidad de la DFT es de O(n2) operaciones y con objeto de acelerar el cálculo de este procedimiento, se emplea habitualmente lo que se conoce como transformada rápida de Fourier o FFT *(Fast Fourier Transform).*

3.2.3 Análisis por Predicción Lineal (LPC)

El método de predicción lineal o LP (*Linear Prediction*) es históricamente uno de los métodos más importantes para el análisis de la voz.

Se fundamenta en el establecimiento de un modelo de filtro del tipo todo polo, para la fuente de sonido.

La principal motivación del modelo todo polo viene dada porque permite describir la función de transferencia de un tubo, que sin pérdidas estuviese formado por diferentes secciones.

Constituye una aproximación razonable al habla producida por la excitación del tracto vocal causada por el conjunto de pulsos glotales.

Con un número suficiente de parámetros el modelo de predicción lineal puede constituir una aproximación adecuada a la estructura espectral de todo tipo de sonidos.

El principio básico del método de predicción lineal recae en el hecho de que en un sistema resonante la salida actual del sistema depende de su salida anterior. Pero esta no es completamente cierta cuando se considera una señal de voz, debido a que la resonancia existente en el tracto vocal varia lentamente con el tiempo, por lo que debería sumarse un error a la salida y estos coeficientes tienen que ser estimados en el tiempo[3].

La ecuación quedaría de la siguiente forma:

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + e(n)$$

[3] VELÁSQUEZ SILVA, Juan Domingo. "*Análisis fino del tracto vocal basado en filtros LPC aplicado al mejoramiento de la calidad de síntesis de voz*". Universidad de Chile, Departamento de Ciencias de la Computación. 1996.

El error de predicción (también conocido como señal residual), e(n), es simplemente la diferencia entre el valor actual de la señal, s(n), y el valor que se predijo, s(n):

$$e(n) = S(n) - \hat{S}(n)a_k$$

Los factores que otorgan el peso, a_k, son encontrados al minimizar el error cuadrático medio, encontrado en N muestras (E):

$$E = \frac{\sum_{i=0}^{N-1} e^2(i)}{2}$$

Los coeficientes a_k que minimizan el error de predicción En son calculados igualando el gradiente con respecto a a_i a 0, para i = 1, .., k. Lo que da como resultado una serie de ecuaciones lineales:

$$\frac{\partial E}{\partial a_k} = 0 \qquad \forall k$$

Se sabe que s(n) es constante, porque es la señal original, luego al derivar para encontrar el mínimo se tiene que:

$$\frac{\partial E}{\partial a_k} = \sum_i \left(S(i) - \sum_j a_j S(i-j) S(i-k) \right) = 0 \qquad j = 1...p$$

$$\sum_i \left(S(i) S(i-k) \right) = \sum_j a_j \left(\sum_i S(i-j) S(i-k) \right)$$

Si se define $\gamma(i,k) = \sum_n S(n-i) \bullet S(n-k)$ se obtiene un sistema de ecuaciones matriciales de la forma:

$$\begin{matrix} \gamma_{(1,1)} & \cdots & \gamma_{(1,1)} & a_1 \\ \cdots & \cdots & \cdots & \cdots \\ \gamma_{(p,1)} & \cdots & \gamma_{(p,\,p)} & a_p \end{matrix} = \begin{matrix} \gamma_{(1,1)} \\ \cdots \\ \gamma_{(p\,,\,0)} \end{matrix}$$

Pero se tiene un problema, para medir en los bordes de la ventana de la señal se necesita salir de la ventana, por lo tanto se pueden hacer 2 suposiciones:

1. Medir fuera de la ventana suponiendo que es cíclica.

2. Todo fuera de la ventana es 0, lo cual otorga una nueva función g cuyos límites son -. y +., esto simplifica bastante el problema, llegando al siguiente estado:

$$\gamma_{(i,k)} = r_{|i-k|} = \sum_{n=-\infty}^{+\infty} S(n)S(n-(i-k))$$

La ecuación anterior bajo esas condiciones es conocida como método de auto-correlación. Esto es casi equivalente a suponer que la señal se repite y que la función fuera de la ventana vale 0, por lo tanto la matriz adopta la siguiente forma:

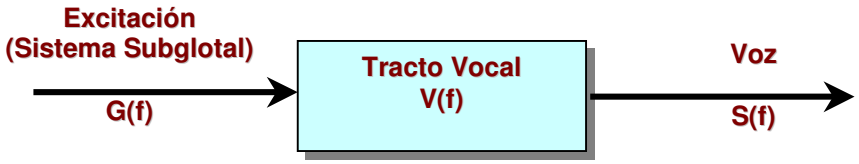
$$\begin{matrix} r_0 & \cdots & r_{p-1} & a_1 \\ \cdots & \cdots & \cdots & \cdots \\ r_0 & \cdots & r_{p-1} & a_1 \end{matrix} = \begin{matrix} r_1 \\ \cdots \\ r_p \end{matrix}$$

3.2.4 Coeficientes Cepstrales

Desde la introducción en los primeros años de la década de los 70, de las técnicas homomórficas de procesado de señal, su importancia dentro del campo del reconocimiento de voz ha sido muy grande.

Los sistemas homomórficos son una clase de sistemas no lineales que obedecen a un principio de superposición. De éstos, los sistemas lineales constituyen un caso especial.

La motivación para realizar un procesado homomórfico del habla viene resumida en la Figura 3.4.



Las técnicas homomórficas sirven para separar la acción del tracto vocal de la señal en el tiempo

El análisis Cepstral se basa en las propiedades homomórficas de funciones logarítmicas:

$$S(f) = G(f)V(f)$$

$$\log[S(f)] = \log[G(f)V(f)]$$

$$\log[S(f)] = \log[G(f)] + \log[V(f)]$$

Donde al menos en teoría, la función del tracto vocal y el espectro de excitación puede ser separados por un filtro pasa bajos.

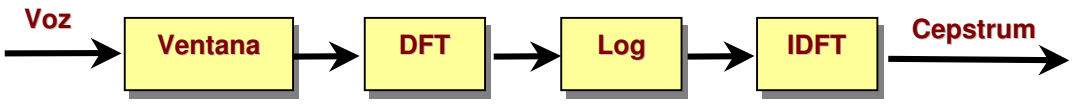


Figura 3.5. Análisis cepstral partiendo de la Transformada Discreta de Fourier.

$$c(n) = \frac{1}{N_S} \sum_{k=0}^{N_S-1} \log_{10} \left| S_{med}(k) \right| e^{j\frac{2\pi}{N_S}kn} \qquad 0 \le n \le N_S - 1$$

Es este caso, el valor *c(n)* se conoce como coeficientes cepstrales derivados de la transformada de Fourier. *Ns* es el número de puntos con los que se calculó la *DFT*.

Esta ecuación también se conoce como la inversa de la DFT del espectro logarítmico. Puede ser convenientemente simplificada teniendo en cuenta que el espectro logaritmo es una función real simétrica y por transformaciones del coseno:

$$c(n) = \frac{2}{N_S} \sum_{k=1}^{N_S} \left(I(k) \right) \cos\left(\frac{2\pi}{N_S} kn \right)$$

Lo habitual es usar solamente los primeros términos ($n \le 20$). *I(k)* representa una función que traduce la posición de un valor en frecuencia al intervalo donde esté contenido.

Es posible a la hora de calcular un coeficiente cepstral, emplear bandas definidas según escalas de Mel.

Este tipo de parámetros se conoce como coeficientes cepstrales con frecuencia en escalas de Mel o MFCC (*Mel Frequency Cepstral Coefficients*).

3.2.4.1 Mel Frequency Cepstral Coefficients. Una representación compacta de la señal acústica puede ser proporcionada por un conjunto de coeficientes Cepstrum en la escala Mel (MFCC), resultados de transformaciones coseno del logaritmo del espectro de energía expresado en la escala de Mel-frecuencia. Los MFCP han demostrado ser los más eficientes en la extracción de parámetros de la señal de voz. El cálculo del MFCC se hace de la siguiente forma:

1. La DFT transforma el segmento de habla ventaneado en el dominio de frecuencia donde se obtiene el espectro de energía P (f).

2. El espectro P (f) se alinea a lo largo de su eje de frecuencia f (en hertzio) en el eje de Mel-frecuencia, como P(m) donde m es la Mel-frecuencia, usando la siguiente ecuación. Esta es una aproximación de la percepción en el oído humano:

$$Escala\ Mel = 2595 \log_{10}\left(\frac{1+f}{700}\right)$$

3. El espectro de potencia resultante se convoluciona con el filtro triangular pasa-banda P(m) en $\theta(m)$. La convolución con el espectro de potencia de banda crítica relativamente ancho enmascarando curvas $\Psi(m)$ que significativamente reducen la resolución espectral de $\theta(m)$, comparado con el espectro original P(f), la convolución discreta de $\Psi(m)$ con $\theta(m)$ proporciona muestras banda critica del espectro de potencia como $\theta(m_k)$, (k=1 ..K) en la siguiente ecuación:

$$\theta(M_k) = \sum_M P(M - M_k)\Psi(m), \qquad k = 1...K$$

Con esto se obtienen las K salidas donde X(k)=ln(θ(m_k)) (k=1...K). La MFCC es calculada usando la siguiente ecuación donde D<<K, debido a la capacidad de compresión de MFCC[4].

$$MFCC(d) = \sum_{k=1}^{L} X_k COS\left[d(k-0.5)\frac{\pi}{L} \right] \qquad d = 1...D$$

Donde: k es la banda de frecuencias.
X_k es el logaritmo de la suma de módulos de la FFT en la banda k.
L es el número de bandas o filtros.
d es el número de coeficiente que se calcula.
D es el número total de coeficientes MEL.

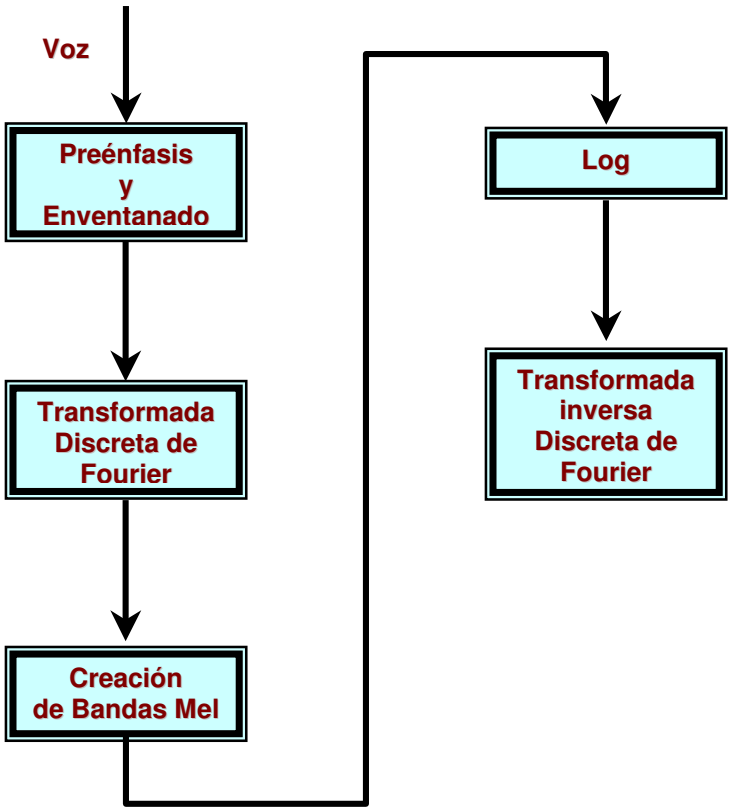


Figura 3.6. Esquema de parametrización para la obtención de MFCC.

[4] PICONE, J.W. , "Signal modeling techniques in speech recognition," *IEEE*, 1993. p. 4 .

3.2.5 Predicción lineal Perceptual.

La técnica de predicción lineal perceptual o PLP (*Perceptual Linear Prediction*), es en esencia una combinación de las técnicas de la transformada discreta de Fourier y de predicción lineal como puede verse en la Figura 3.6.

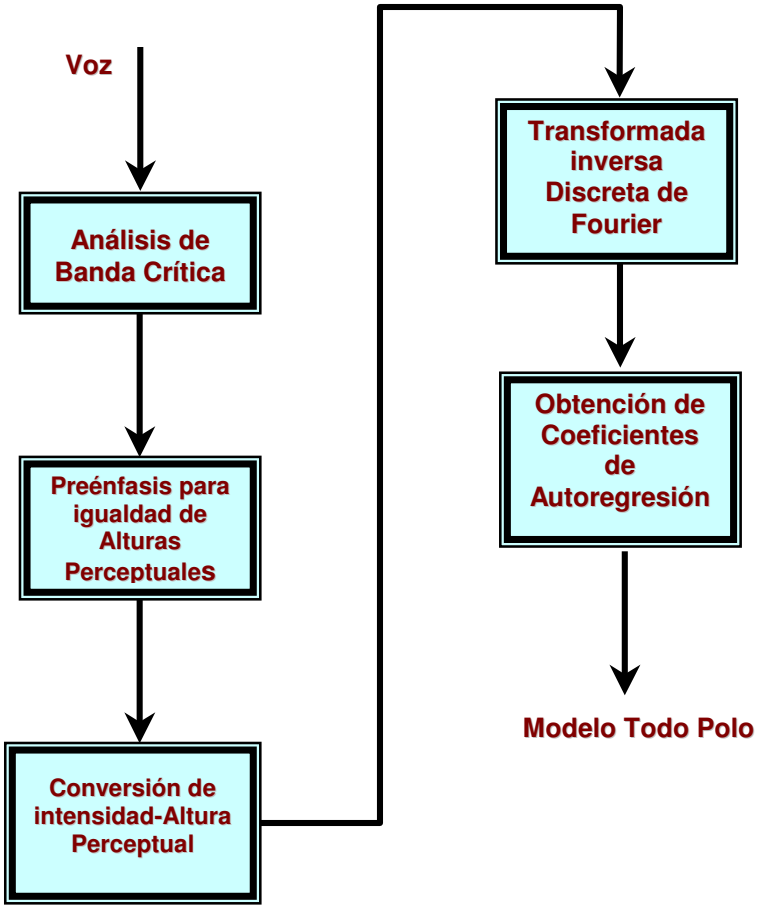


Figura 3.6. Predicción Lineal Perceptual

Para obtener el análisis de banda crítica se utiliza primeramente la transformada discreta de Fourier con una ventana de Hamming de 20 ms.

Posteriormente se calcula el espectro de potencia y se transfiere a una escala de Bark.

El segundo paso consiste en la igualación de las alturas perceptuales tiene su origen en la necesidad de compensar la diferente percepción de alturas sonoras para diferentes frecuencias.

Tras la IDFT se calculan los coeficientes de autorregresión de un modelo todo polo.

Adicionalmente se pueden calcular a partir de éstos los coeficientes cepstrales. Como extensión de la técnica anteriormente descrita, se encuentra el método RASTA-PLP (*Relative Spectral*).

La motivación de este complemento viene dada por el intento de robustecer el mecanismo del algoritmo frente a distorsiones lineales en el espectro, por ejemplo debidas al canal de comunicación.

Una extensión del algoritmo RASTA es la conocida como J-RASTA, que puede también compensar el ruido cuando la relación señal / ruido es baja.

4. TÉCNICAS CLÁSICAS APLICADAS AL RECONOCIMIENTO DE HABLA

4.2 ALINEAMIENTO TEMPORAL

El método de alineamiento temporal o *DTW* (*Dynamic Time Warping*) es uno de los algoritmos aplicados en reconocimiento de voz más antiguos e importantes.

En la actualidad ha cedido paso a otros procedimientos como son los modelos ocultos de Markov que también se estudian en este capitulo.

Si bien esta técnica aún se continúa utilizando, tiene un número de limitaciones que restringen su uso a sistemas con vocabularios pequeños.

En sistemas de mayor tamaño, el número de plantillas a generar y el coste computacional de las búsquedas es intratable.

La manera más fácil de reconocer una palabra aislada es compararla con un conjunto de plantillas previamente almacenadas y determinar cual es la que proporciona un mejor encaje.

Sin embargo, este objetivo se complica por dos factores:

- La duración de la palabra no tiene que ser la misma que la de las plantillas.
- El ritmo con el que pronuncia esa palabra no tiene por que ser constante.

Resumiendo, el alineamiento óptimo, entre plantillas almacenadas y las plantillas producidas en un determinado momento, puede ser no lineal.

La técnica de Alineamiento Temporal (*Dynamic Time Warping*) se encarga de realizar la comparación de patrones acústicos de la señal de voz. En ella se

tiene en cuenta la variación en la escala del tiempo de dos palabras a comparar.

Una pronunciación queda representada en esta aproximación mediante una secuencia de longitud variable de vectores de características (Coefientes Ceptrales Espectro LPC etc.). Cada vector de características es el resultado de aplicar algún tipo de transformación sobre la señal vocal a través de una ventana de análisis de tamaño fijo que se desliza sobre la pronunciación en intervalos de tiempo (mediante técnicas estudias en el capitulo anterior).

Dado un conjunto de muestras de entrenamiento y sus correspondientes transcripciones fonéticas la técnica empleada permite segmentar cada muestra de entrenamiento en unidades de tipo fonético mediante un algoritmo convencional de alineación temporal no lineal, obtenido, una vez segmentadas todas las muestras, un conjunto de plantillas por cada unidad fonética, donde cada plantilla es una secuencia de vectores de características.

Sean $X= (x_1, x_2, ..., x_I)$ e $Y= (y_1, y_2, ..., y_J)$ dos patrones de voz aislados (ej. palabras). La disparidad promedio entre X e Y, $D(X,Y)$, se basa en alguna medida de distancia entre los vectores x_i e y_j que denotaremos como $d(i,j)$.

La distorsión temporal no lineal de los patrones X e Y puede representarse por un camino $\{ P(k)= (m(k), n(k)), k= 1, K \}$ en el plano (i, j) definida por las dos secuencias de vectores X e Y.

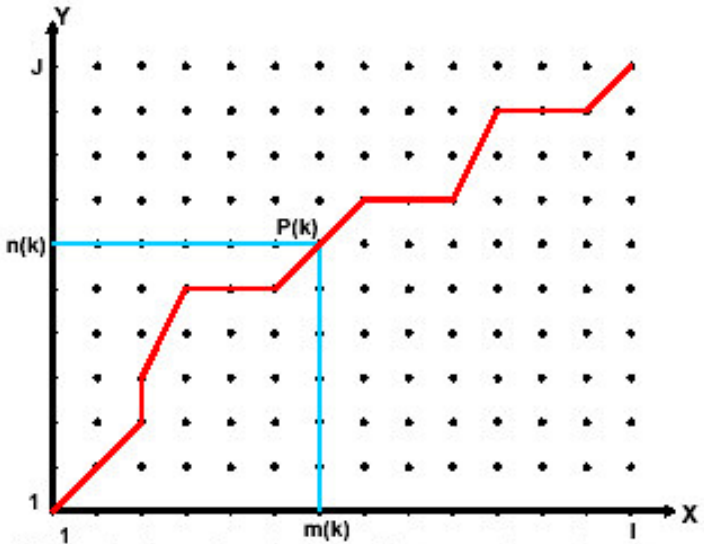


Figura 4.1. Ejemplo de camino de Alineamiento temporal para los patrones de voz X e Y

La disparidad entre las tramas *X* e *Y* a lo largo de un camino *P* viene dada por:

$$D_p(x, y) = \sum_{k=1}^{K} d(P(k)) \frac{w(k)}{N(w)}$$

w(k) es un peso y *N(w)* es el factor de normalización. Existe un número de posibles caminos *P(k)*, que corresponden a diferentes funciones de distorsión para los patrones de voz[5].

El objetivo será encontrar el camino que minimice $D_p(X,Y)$, siendo una elección natural el tomar el mínimo sobre todos los posibles caminos:

$$D(X, Y) = \min_P (D_P(X, Y))$$

Para resolver este problema se emplean técnicas inspiradas en los algoritmos de programación dinámica.

---

[5] VIDAL, José E. Decodificación Acústico-Fonética mediante plantillas Subléxicas. España: Comisión Internacional de Ciencia y Tecnología, 1990 p. 1-2

Con objeto de tener en cuenta ciertos aspectos físicos del problema y limitar el número de caminos a considerar, se imponen algunas restricciones a las funciones de distorsión:

- Puntos de inicio y final de tramas. P(1) = (1,1) ; P(K)=(I , J)
- El camino no podrá tener una pendiente negativa.
- Continuidad Local. Para minimizar la pérdida de información se restringen los movimientos locales.
- Restricciones de pendiente. Se expresan como funciones de costo para el cálculo del peso w(k).

Por su parte, el factor de normalización suele ser:

$$N(w) = I + J$$

Una variación importante del DTW es una extensión de reconocimiento de palabras aisladas a habla continua. Esta extensión es llamada algoritmo DTW de un Paso. En este caso particular el objetivo es encontrar el alineamiento optimo entre la muestra de habla y la mejor secuencia de palabras de referencia (ver Figura 4.2).

4.2.1 Algoritmo de un paso

Un enfoque habitual para encontrar la mejor hipótesis de frase de una muestra de habla continua es el algoritmo de un paso *(One-Stage-Dynamic-Time-Warping),* recogido por Sakoe (1979) y Ney (1984).

Se trata de un algoritmo Viterbi generalizado que combina la segmentación con el proceso de reconocimiento. En el seno de la cadena Markov, y para una palabra, el algoritmo es capaz de encontrar la probabilidad para el mejor camino a través de la palabra empleando el algoritmo Viterbi.

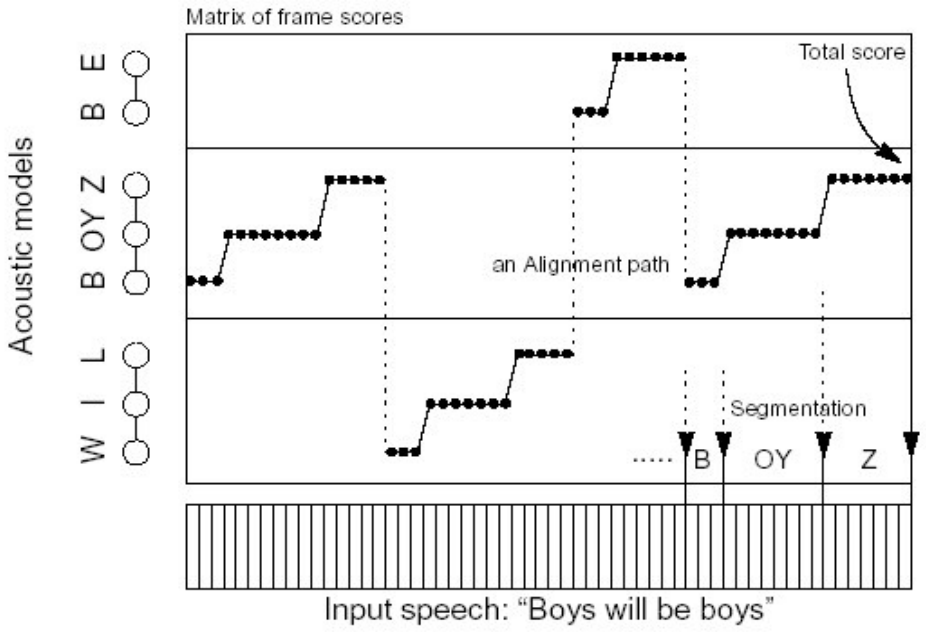


Figura 4.2. El camino de alineación con los mejores *scores*, la secuencia de palabras y su segmentación

Por lo que, para cada *frame* existen transiciones adicionales posibles desde todos los finales de palabras a todos los estados de inicio de una nueva palabra. Mientras en el algoritmo Viterbi normal el único precedente legal de un estado de inicio de palabra es ese el mismo estado de inicio de palabra, todos los estados de final de palabra se consideran ahora posibles predecesores.

Dado que las multiplicaciones requeridas pueden ser reemplazadas por sumas en el espacio logarítmico es habitual utilizar el logaritmo negativo de las probabilidades acumulativas. En consecuencia, una puntuación alta implica una baja probabilidad y una puntuación baja, una alta probabilidad. Para reconstruir la mejor secuencia de palabras, la única información requerida es qué palabra es el mejor precedente al final de la palabra en curso. Así, para cada final de palabra, se almacena la siguiente información: cuál era el mejor antecesor de la

palabra y en qué *frame* se dio la transición desde el predecesor a la palabra actual. Esta estructura se denomina *backtrace (rastreo hacia atrás)*[6]

Si no se utiliza ningún modelo de lenguaje, el mejor predecesor de una palabra para todas las palabras que empiezan en un *frame* es la palabra con la mejor puntuación en su estado de final de palabra para el *frame* anterior. En ese caso, el *backtrace* tiene el mismo aspecto para todas las palabras y podría ser reemplazado por un simple indicador por *frame*. Si, pese a ello, el mejor predecesor depende de la identidad de la palabra debido a un modelo de lenguaje de *dos letras (bigram)*, hay que almacenar un *indicador retrospectivo (backpointer)* por cada *frame* y por cada final de palabra.

La información que debe almacenarse en el backtracec hasta alcanzar el último estado de una palabra podría obtenerse siguiendo los punteros situados en esa palabra de estado a estado. Sin embargo, para el *frame* en curso resulta mucho más eficiente mantener la información de la palabra de entrada al *frame* y la palabra que la precede en la estructura de la información de cada estado. En el siguiente *frame*, cada estado en el seno de la palabra hereda esta información de su mejor estado predecesor. Así, no es necesario almacenar más información sobre la ruta exacta entre las palabras. Los requisitos de memoria se reducen a la empleada para el rastreo hacia atrás, más los contenidos de todos los estados del *frame* en curso y el previo.

4.3 MODELOS OCULTOS DE MARKOV

La técnica más flexible utilizada para el reconocimiento de habla hasta ahora ha sido los Modelos Ocultos de Markov (HMMs de sus siglas en inglés).

---

[6] L. R. Rabiner y S. E. Levinson, "*Isolated and Connected Word Recognition- Theory and Selected Applications*", IEEE Tratados sobre Comunicaciones, Vol. COM-29, Nº. 5, 1981, pp. 121-129.

En esta sección se exponen los conceptos básicos de los HMMs, se describen los algoritmos de entrenamiento y utilización de estos, las variaciones más comunes y además se destacan algunos problemas asociados con ellos.

4.3.1 Conceptos Básicos

Considerando que el *frame* en el instante *t* se compone por un vector de parámetros espectrales O_t, una elocución se representa entonces por una secuencia de vectores *O*:

$$O = \left[o_1, o_{2,} o_3, ..., o_T\right]$$

donde *T* es la duración en *frames* de la señal. Para realizar la clasificación de patrones acústicos en el proceso de VL se debe medir la "distancia" entre el modelo del habla del locutor afirmado *j* y la secuencia de vectores de observación *O* del locutor *i* que clama dicha identidad. Luego, la distancia obtenida es comparada con un límite de decisión, con el cual se aceptará o rechazará la identidad clamada. En el caso de los HMM, la distancia corresponde a una probabilidad, la cual se define como:

$$P(S_i = S_j / O, \lambda_j)$$

donde S_i corresponde al locutor *i* que pretende ingresar al sistema, S_j corresponde al cliente *j* que dice ser, *O* el vector de observación obtenido de la elocución de verificación, y λ_j el modelo de referencia del cliente *j* generado a partir de las elocuciones de entrenamiento. Usando el Teorema de Bayes para probabilidades condicionadas se tiene:

$$P(S_i = S_j / O, \lambda_j) = \frac{P(O / S_i = S_j, \lambda_j) \cdot P(S_i = S_j)}{P(O)}$$

Debido a que los términos $P(S_i = S_j)$ y $P(O)$ se consideran constantes para todos los locutores, el término trascendental para encontrar $P(S_i = S_j / O, \lambda_j)$ corresponde al valor de la verosimilitud definida por $P(O / S_i = S_j, \lambda_j)$.

4.3.2 Clasificación de Patrones usando HMM

Para los procesos de clasificación mediante modelos ocultos de Markov (HMM) se asume que cada secuencia de vectores de observación corresponde a palabras de un locutor determinado y que cada una de ellas es generada por un modelo de Markov.

Un modelo de Markov consiste en una secuencia finita de estados conectados entre sí por probabilidades de transición. Cada unidad temporal, que en este caso corresponde al *frame*, debe enfrentarse ante la posibilidad de mantenerse en el estado actual o avanzar al siguiente.

Cada estado *x* se caracteriza por una función de densidad de probabilidad de observar un cierto *frame* O_t. Esta función también se denomina probabilidad de salida o de emisión y se denota por $b_x(o_t$ Considerando la topología izquierda-derecha sin salto de estados de la Fig. 4.3, la probabilidad de transición desde

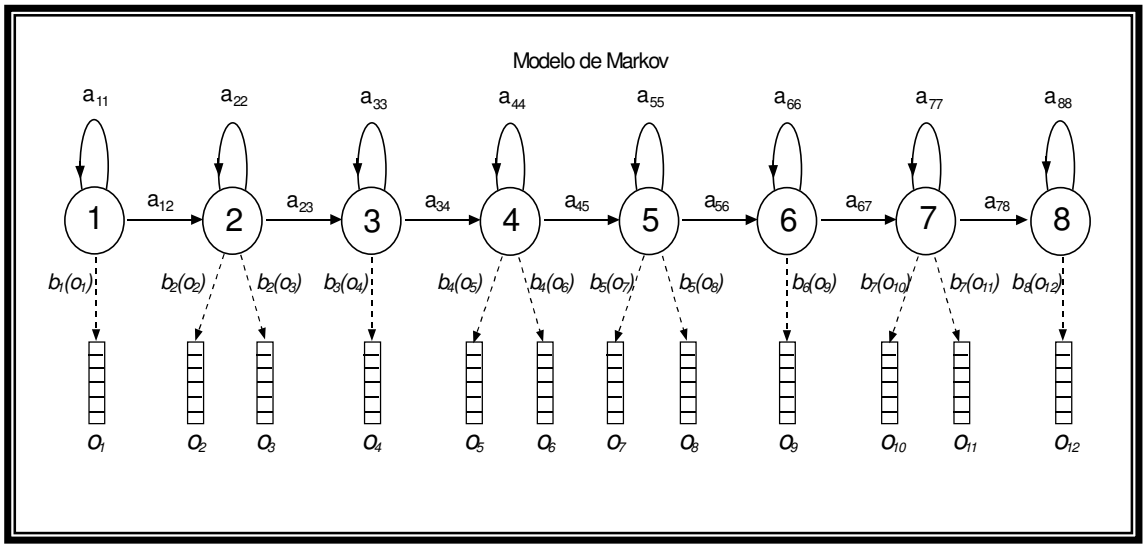


Figura 4.3. Topología izquierda derecha sin salto de estado de un HMM

El estado *x* al estado *y* es $a_{x,y}$, donde *y=x* o *y=x+1*. Por definición se tiene que $a_{x,x} + a_{x,x+1} = 1$.

Con las definiciones discutidas hasta ahora se tiene que el modelo de referencia es $\lambda_j = (A, B, \pi)$ donde A es la matriz de todas las transiciones de probabilidad, B es el conjunto de los parámetros de las probabilidades de observación, y π son las probabilidades de que cada estado sea el primero.

En el modelo de Markov descrito en la Fig. 3 se puede identificar la secuencia de estados como: $X = \{1,2,2,3,4,4,5,5,6,7,7,8\}$ generada por la secuencia de observación *O*=[o_1, o_2, o_3, ... o_{12}].

La probabilidad conjunta de que el vector de observación *O* sea generado por el modelo λ_j de la identidad clamada moviéndose a través de la secuencia *X*, es calculada como el producto entre las probabilidades de transición y las probabilidades de observación. De esta forma, para la secuencia *X* mostrada en la Fig. 4.3 se tendrá:

$$P(O, X / \lambda_j) = b_1(o_1) \cdot a_{12} b_2(o_2) \cdot a_{22} b_2(o_3) \cdot a_{23} b_3(o_4) \cdot a_{34} b_4(o_5) \cdot a_{44} b_4(o_6) \cdots\cdots$$

En la práctica, sólo la secuencia de observación *O* es conocida y la secuencia de estados *X* correspondiente es oculta. Este es el motivo por el cual estos modelos se denominan Modelos Ocultos de Markov. De esta forma, dado que *X* es desconocido, la verosimilitud requerida es computada mediante la sumatoria de todas las posibles secuencias de estados *X = {x(1), x(2), x(3),.....,x(T)}*,

$$P(O / \lambda_j) = \sum_{\text{todos } X} \prod_{t=1}^{T} b_{x(t)}(o_t) \cdot a_{x(t-1)x(t)}$$

donde *x(0)*=1 corresponde al estado inicial del modelo de la Fig. 4.3. Una aproximación para la verosimilitud, consiste en considerar solamente la secuencia de estados más probable:

$$P(O / \lambda_j) \cong \underset{X}{Max} \left\{ \prod_{t=1}^{T} b_{x(t)}(o_t) \cdot a_{x(t-1)x(t)} \right\}$$

Si bien el cálculo analítico de la ecuación 10 no es posible, existen procedimientos recursivos que permiten calcular esta expresión de manera eficiente. Uno de estos procedimientos corresponde al algoritmo de Viterbi, el cual determina una secuencia de estados óptima y la respectiva verosimilitud. Por su parte, las matrices *A* y *B* son determinadas con las elocuciones de entrenamiento del sistema utilizando el algoritmo de re-estimación de Baum-Welch (Deller et al., 1993)[7].

[7] A. B. Poritz, "*Hidden Markov Models: A Guided Tour*", ICASSP'88, Nueva York, Estados Unidos: 1988, p. 7-13.

4.3.3 La Probabilidad de Observación ($b_j(o_t)$).

Los parámetros de los vectores de observación (o_t) asumen valores continuos y la probabilidad de observación se puede modelar con una función de densidad de probabilidad multivariable. Esta función de densidad de probabilidad está constituida generalmente por una combinación lineal de Gaussianas:

$$b_x(o_t) = \sum_{g=1}^{G} c_{xg}\, \aleph(o_t; \mu_{xg}, \Sigma_{xg}) \qquad ,1 \le x \le N_e$$

donde N_e corresponde al número de estados del HMM, G es el número de Gaussianas, $c_{x,g}$ es la ponderación de las Gaussianas, las cuales deben cumplir:

$$\sum_{g=1}^{G} c_{xg} = 1 \qquad ,1 \le x \le \mathrm{N_e}$$
$$c_{xg} \ge 0 \qquad ,1 \le x \le \mathrm{N_e} \wedge 1 \le g \le G$$

$\aleph(.;\mu,\Sigma)$ corresponde a una Gaussiana multivariable con vector de medias μ y una matriz de covarianza Σ:

$$\aleph(o_t; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n \cdot |\Sigma|}} \cdot e^{-\frac{1}{2}(o_t-\mu)^T \Sigma^{-1} (o_t-\mu)}$$

4.3.4 Algoritmos

4.3.4.1 El Algoritmo de Viterbi

En la sección 4.3.3 se mostró que la probabilidad conjunta de que el vector de observación O sea generado por el modelo λ_j de la identidad clamada

moviéndose a través de la secuencia de estados *X* (o verosimilitud $P(O,X/\lambda_j)$ es calculada como un producto entre las probabilidades de transición y la probabilidad de observación.

Dado que la secuencia de estados *X* no es conocida se debe calcular la secuencia más probable, tal como se mostró en la ecuación 10. Para encontrar la secuencia más probable, y por ende realizar la verificación más eficientemente, se utiliza el algoritmo de decodificación de Viterbi.

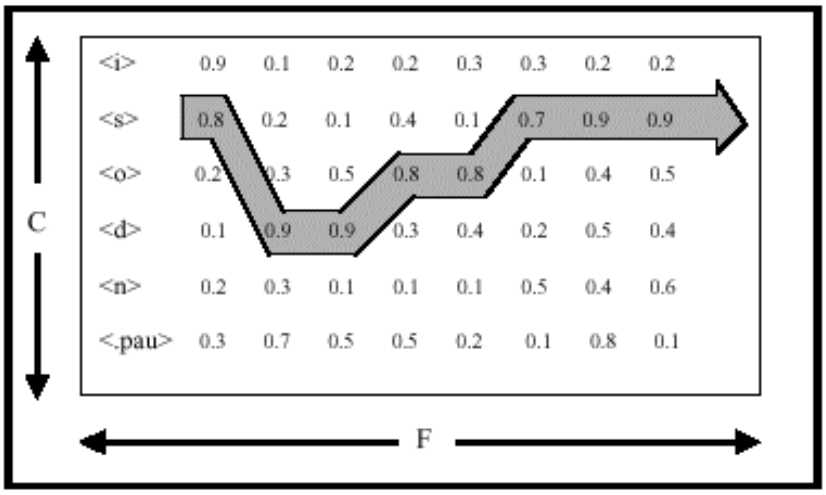


Figura 4.4. Representación gráfica del algoritmo de Viterbi

En la Figura 4.4 se muestra la gráfica que representa al uso del algoritmo de Viterbi operando sobre un modelo HMM de 6 estados con topología izquierda – derecha y sin salto de estado.

El algoritmo de Viterbi puede visualizarse como una solución para encontrar el camino óptimo a través de una matriz que posee como eje vertical los estados del modelo HMM y como eje horizontal los *frames* en los que está dividido la señal de voz. Cada espacio de la matriz mostrado en la figura representa el

logaritmo de la probabilidad de observar aquel *frame*[8]. Si $\hat{X}$ = *{x(1), x(2), x(3),.....,x(t)}* es la secuencia óptima de estados obtenida para la secuencia de vectores de observación *O = [o1, o2,....., ot]*, y además se considera $\delta_y(t)$ como la máxima probabilidad calculada a lo largo de un camino, trazado hasta el tiempo *t*, y finalizada en el estado *y*, se tiene que:

$$\delta_y(t) = P(x(1), x(2),....., x_{(t-1)}, x(t) = y, [o_1 o_2 ... o_t] / \lambda_c)$$

es máxima en función de las posibles secuencias de estados hasta *t*. Luego, la verosimilitud para el instante *t+1* se calculará usando los valores obtenidos de la:

$$\delta_x(t+1) = \underset{y}{Máx} \left(\delta_y(t) \cdot a_{y,x}\right) \cdot b_x(o_{t+1})$$

4.3.5 Normalización de la Verosimilitud

Considerando un enfoque clásico, en los sistemas de VL las decisiones son tomadas calculando la verosimilitud de la elocución de verificación en relación al modelo HMM de la identidad clamada. En este caso, el valor de la verosimilitud también considera la información lingüística de la señal de testeo y del modelo del locutor clamado. De esta forma, al valor de la distancia deseada es adicionada una cantidad que depende de la variabilidad natural del habla por lo que un umbral de decisión estándar es difícil de fijar. Para reducir el problema de la variación del umbral de decisión se aplica la normalización de la verosimilitud (Higgins, Bahler & Porter, 1991; Rosenberg, 1992; Matsui & Furui, 1993). Esta normalización mejora significativamente el desempeño de los sistemas de reconocimiento de Habla y se realiza ocupando la relación entre las verosimilitudes de la elocución de test frente al modelo de referencia y con respecto a un modelo global:

[8] Ibid., p. 15-16

$$L(O) = \frac{P(O/S_c)}{P(O/S_{\bar{c}})}$$

La probabilidad de que la secuencia de vectores de observación *O* corresponda al modelo de referencia del locutor *c* $(P(O/S_c))$ es calculada como se mostró hasta ahora. Por su parte, el $P(O/S_{\bar{c}})$, denominado término normalizador, corresponde a la verosimilitud calculada con respecto a un modelo general de *posteriors*.

5. TÉCNICAS CONEXIONISTAS APLICADAS A RECONOCIMIENTO DE HABLA

5.1 REDES NEURONALES ARTIFICIALES

Existen muchos tipos de redes neuronales, pero todas estas presentan cuatro atributos básicos que son:

- Un set de Unidades de procesamiento;
- Un set de conexiones;
- Procesos de computo;
- Procesos de entrenamiento.

5.1.1 Unidades de Procesamiento

Una red neuronal contiene un gran número de unidades de procesamiento simple, análogo a las neuronas del cerebro. Todas estas unidades operan simultáneamente, soportando el paralelismo masivo. Todos los cálculos en el sistema son hechos por estas unidades. No hay otro procesador que coordine sus actividades. En cada momento cada unidad simplemente computa una función escalar de sus entradas locales y evalúa el (llamado valor de activación) a las unidades adyacentes[9].

[9] HILERA, José R. y MARTÍNEZ, Victor J. Redes Neuronales Artificiales. Madrid, España: Alfa-Omega Editor, 2000. p. 51 - 56

Las unidades en una red se dividen en: unidades de entrada, las cuales reciben los datos del exterior (parámetros de la señal de voz en este caso); unidades ocultas, las cuales pueden transformar internamente los datos y unidades de salida, las cuales representan decisiones o señales de control.

En los esquemas de redes neuronales, las unidades son representadas por círculos. También, por convección, las unidades de entrada se muestran debajo mientras que las de salida se muestran en la parte superior del esquema.

Una unidad de proceso recibe varias entradas procedentes de las salidas de otras unidades de proceso. La entrada total de una unidad de procesos se suele calcular como la suma de todas las entradas ponderadas, es decir, multiplicadas por el peso de la conexión. El efecto inhibitorio o excitatorio de la sinapsis se logra usando pesos negativos y positivos respectivamente.

5.1.2 Conexiones

Las unidades se encuentran organizadas en la red, mediante una topología dada, por un set de conexiones, o pesos, que se muestran en el diagrama como líneas de conexión. Cada peso tiene un valor real, típicamente en el rango de $-\alpha$ a $-\alpha$ aunque algunas veces este rango es limitado, el valor de un peso describe cuanta influencia tiene una unidad sobre sus unidades adyacentes; Un peso positivo ocasiona que una unidad excite a otra, mientras que un peso negativo ocasiona la inhibición de una con respecto a la otra. Los pesos son usualmente unidireccionales (de las entradas hacia las salidas), aunque estos también pueden ser bidireccionales (especialmente cuando no hay distinción entre unidades de entrada o salida)[10].

Los valores de todos los pesos predeterminan la reacción computacional de la red a cualquier parámetro de entrada arbitrario. Los pesos pueden cambiar

[10] Ibid, p. 72.

como resultado del entrenamiento, pero estos tienden a cambiar lentamente, debido a que el aprendizaje acumulado cambia lentamente. Esto contrasta con los patrones de activación, que son funciones transitorias de la entrada actual.

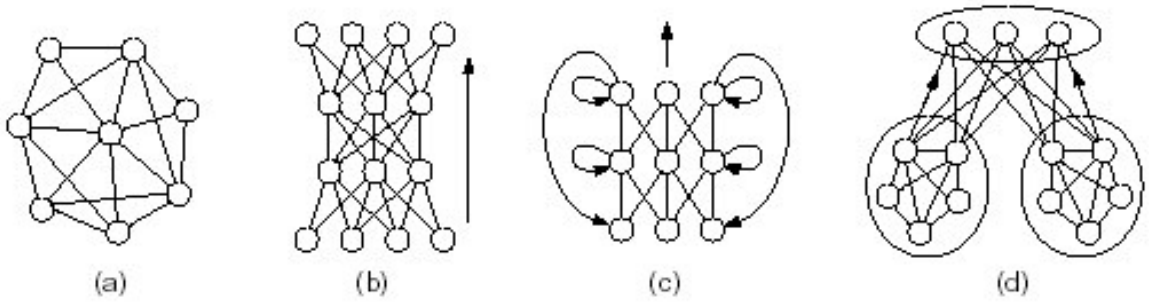


Figura 5.1. Topologías de las redes neuronales: (a) No estructurada, (b) Por capas, (c) Recurrente, (d) Modulares

Las topologías mas comunes se muestran el la Figura 5.1. Cada topología tiene un mejor comportamiento para un tipo de aplicación en particular. por ejemplo:

- La redes no estructuradas son muy utilizadas en la completación de patrones.
- Las redes por capas son utilizadas en la asociación de patrones.
- Las redes recurrentes que se utilizan para el secuenciamiento de patrones.
- Las redes modulares que se utilizan para la construcción de sistemas complejos mediante componentes mas simples.

5.1.3 Computación

La computación siempre comienza, presentando un parámetro de entrada a la red. Luego se computan las activaciones de todas las unidades, ya sea

sincrónicamente (Todas a la vez en un sistema paralelo) o asincrónicamente (una a la vez, un orden natural o aleatorio). En redes no estructuradas, este proceso es llamado extensión de la activación, en redes por capas, este es llamado propagación *forward* (hacia delante), a medida que progresa de la capa de entrada a la capa de salida. En redes *feedfoward*, las activaciones se estabilizarán en el momento en el que la activación alcance la ultima de salida, pero en las redes recurrentes, la activación puede que nunca se estabilice, siguiendo una trayectoria dinámica a través del espacio de estado donde las unidades son actualizadas continuamente.

Una unidad dada es actualizada típicamente en dos pasos: Primero se computan las entrada (o activaciones internas) y luego se computan sus activaciones de salida como una función de la entrada.

En el caso estándar , la entrada de la red x_f para la unidad j es la suma de sus entradas multiplicadas por el respectivo peso:

$$x_f = \sum_i y_i w_{ji}$$

Donde y_t, es la activación de salida de una unidad de entrada, y w_{jt} es el peso de la unidad i a la unidad j

En general, la entrada de red la red es compensada por una ganancia θ por lo que la ecuación para esta es de hecho:

$$x_f = \sum_i y_i w_{ji} + \theta_j$$

Aunque en la practica esta ganancia se trata como si fuera otro peso W_{j0} conectado a una unidad invisible con una activación $y_0=1$

Una vez se ha computado las entradas a las unidades de la red x_f, se computa la activación de la salida y_f como una función de x_f. Esta Función de activación (también llamada función de transferencia) puede ser deterministica o estocástica, y local o no local.

Función de activación deterministica local, que suelen ser de tres formas – *Linear*, *threshold* o *sigmoidal* – como se muestra en la Figura 5.2. En el caso *Linear*, se tiene simplemente que y = x. Esta se utiliza muy poco ya que no es muy potente: múltiples capas de unidades lineales pueden colapsar en una capa simple con la misma funcionalidad. Con el fin de construir funciones no-lineales, una red neuronal requiere de unidades no-lineales. La forma mas simple de no linealidad se proporciona por la función de activación threshold, ilustrada en el panel (b):

$$y = \begin{cases} 0 & si \quad x \leq 0 \\ 1 & si \quad x > 0 \end{cases}$$

Esta es mucho más poderosa que la función lineal, ya que una red multicapa de unidades *threshold* puede teóricamente computar cualquier función booleana. Sin embargo esta es difícil de entrenar debido a las discontinuidades en la función, por lo que para encontrar la configuración deseada de los pesos puede hacerse necesaria una búsqueda exponencial.

Sin embargo existen muchas aplicaciones donde se prefieren salidas continuas en vez de binarias. Consecuentemente, la función mas común es en estos momentos la función *sigmoidal*, ilustrada en el panel (c):

$$y = \frac{1}{1+e^{-x}} \quad \text{o similarmente} \qquad y = \tanh(x)$$

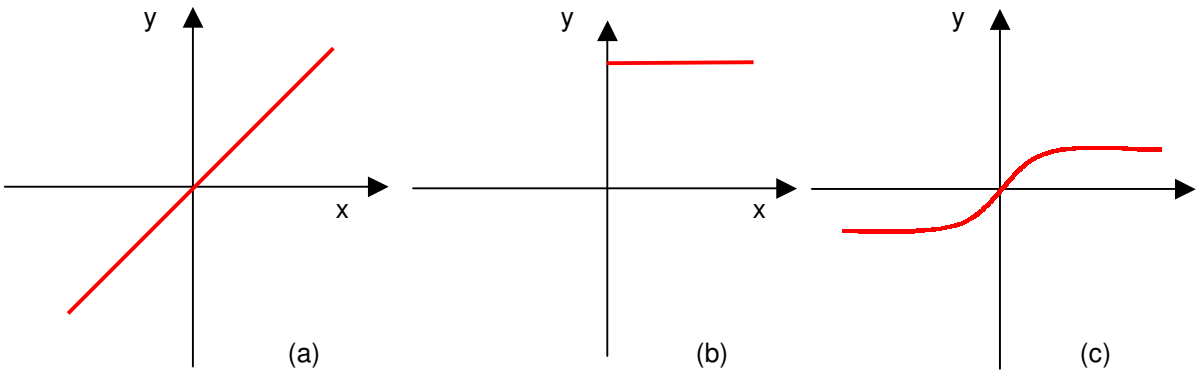


Figura 5.2. Funciones deterministicas de activación local (a) Lineal, (b) threshold, (c) sigmoidal.

Las funciones *sigmoidales* tienen la ventaja de la no linealidad, continuidad, y diferenciabilidad, permitiendo a una red multicapa computar cualquier valor real arbitrario para una función, y además sirve de base para algoritmos de entrenamiento como el de Backpropagation, basado en el gradiente descendiente que se explicara en secciones posteriores.

- Las funciones de activación no locales, se utilizan para imponer limitaciones globales sobre las redes. Por ejemplo, algunas veces se suele forzar todas las activaciones de salida de las redes a que su suma sea igual a 1 (es el caso de las probabilidades). Esto se puede lograr linealmente normalizando las salidas, pero una técnica mas popular es utilizar la función softmax :

$$y_j = \frac{e^{x_j}}{\sum_i e^{x_j}}$$

5.1.4 Entrenamiento

Entrenar una red neuronal, consiste en la adaptación de sus conexiones con el fin de que la red tenga un comportamiento computacional deseado para todo

patrón de entrada. El procedimiento generalmente se realiza modificando los pesos, pero algunas veces también se debe modificar la topología de la red.

En este sentido, la modificación de los pesos es la más común, ya que una red con abundantes conexiones puede aprender a llevar cualquiera de sus pesos a cero, pudiendo resetear los pesos.

El encontrar un set de pesos que le de la capacidad a una red dada de computar una función no es un procedimiento muy trivial.

- Gradiente descendiente

Este algoritmo esta basado en un procedimiento iterativo, que requiere múltiples pasadas de entrenamiento sobre toda la set (de entrenamiento); cada pasada es llamada iteración o *epoch*. Sin embargo, ya que todo el conocimiento acumulado es distribuido sobre todos los pesos, estos deben modificarse levemente, para no destruir todo el aprendizaje previo, por lo que se utiliza una constante pequeña llamada rata de aprendizaje (ε) para controlar la magnitud de modificación de los pesos.

Encontrar el valor optimo de la rata de aprendizaje es muy importante ya que si el valor es demasiado pequeño, el aprendizaje toma mucho tiempo; pero si el valor es demasiado grande, se distorsiona todo el aprendizaje previo. Desafortunadamente no hay método analítico para encontrar la optima rata de aprendizaje; esta es optimizada empíricamente, mediante prueba y error.

Otro algoritmo bastante utilizado para el entrenamiento es la regla Delta, la cual se aplica cuando hay un valor deseado para una unidad de una pareja de estas. Esta regla refuerza la conexión entre dos unidades, si hay una correlación entre la activación de la primera unidad y_t el error relativo para la segunda unidad a un valor deseado (*target*) t_t:

$$\Delta w_{ji} = \varepsilon y_i (t_j - y_j)$$

Esta regla disminuye el error relativo si y_i contribuye a esto, por lo que la red esta preparada para computar una salida y_j cercana a t_f si solamente la primera unidad de activación y_i es conocida durante la prueba.

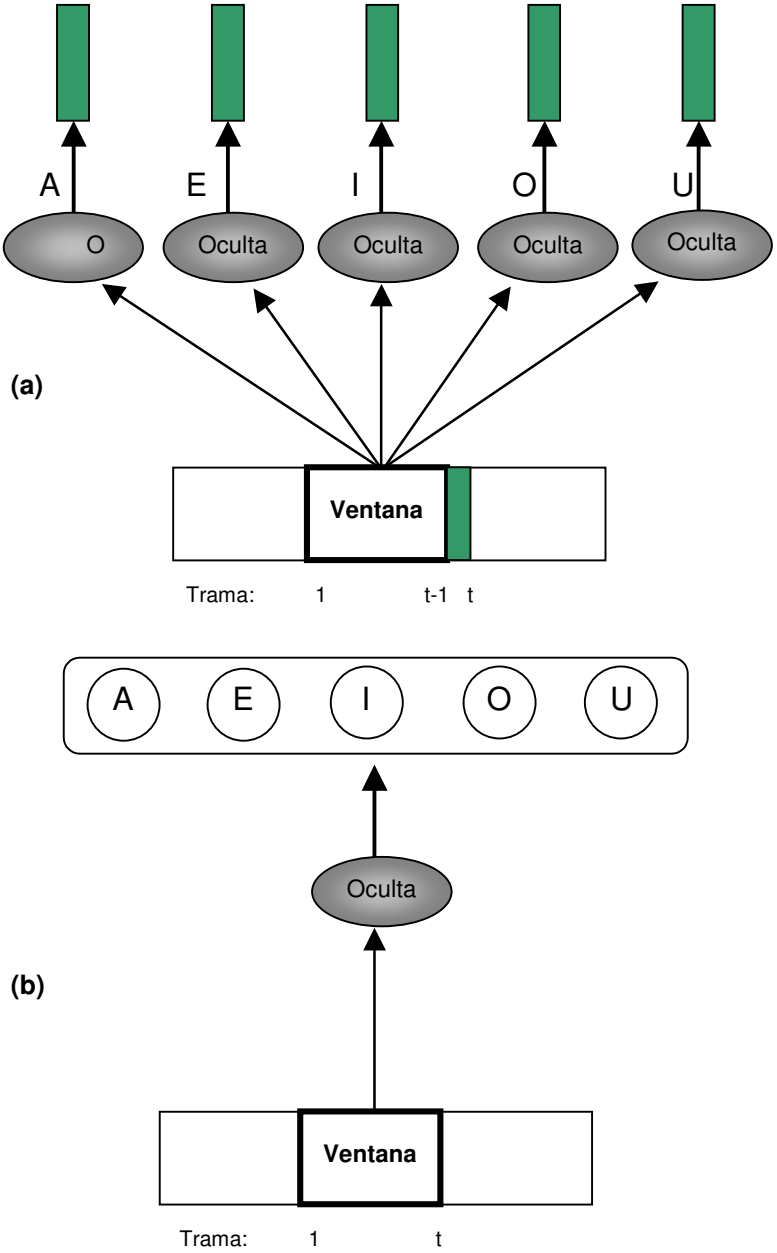
En el contexto de las unidades *threshold* con una capa de pesos, la regla Delta es conocida como la Regla de Aprendizaje Perceptron, y sirve para encontrar el set de pesos representando una solución perfecta, si tal solución existe (Rosenblatt 1962).

En el contexto de redes multicapa, la Regla Delta es básica para el proceso de entrenamiento *backpropagation* que se discutirá en la 5.3.1.3.

5.2 REDES PREDICTIVAS

Las redes neuronales se pueden entrenar para calcular funciones no lineales o no parametrizadas de cualquier espacio de entrada a cualquier espacio de salida.

Dos tipos de funciones muy comunes son la predicción y la clasificación como se muestra en la Figura 5.3



1.1.1 Figura. 5.3 (a) Predicción Vs. (b) Clasificación

En una red predictiva, las entradas son varias tramas de habla, y las salidas son una predicción de la siguiente trama de la señal; utilizando múltiples redes predictivas los errores de predicción pueden ser comparados y la red que tenga el menor error de predicción se considera el mejor modelo (o el elemento más parecido) al segmento de habla que se está comparando.

En contraste en una red de clasificación, las entradas son otra vez múltiples tramas de señal de habla, peor las salidas directamente clasifican el segmento del habla en una clase dada. Figura 5.3.

5.2.1 Redes Neuronales LPNN

La arquitectura Linked Predective Networks (LPNN) está diseñada para el reconocimiento de vocabulario extenso. Tanto para palabras aisladas como para el habla continua están basadas en modelos de fonemas relacionados sobre diferentes contextos.

En esta sección se describen la operación básica y entrenamiento de una LPNN.

5.2.2 Operación Básica.

Una LPNN implementada al reconocimiento de fonemas se muestra en la Figura 5.4. Una red mostrada como un triángulo, toma k tramas continuas de la señal de habla (normalmente se toma k = 2); estas pasan a través de una capa de neuronas ocultas e intenta predecir la siguiente trama de la señal. La trama que se predice es luego comparada con la trama actual. Si el error es pequeño, la red es considerada como un buen modelo para ese segmento de la señal. Si

se pudiera enseñar a la red a hacer predicciones exactas solamente durante los segmentos correspondientes al fonema $|A|$ (en este caso) y predicciones vagas a los demás, se tendrá un reconocedor bastante eficiente del fonema $|A|$, en virtud de su contraste con otros modelos de fonemas.

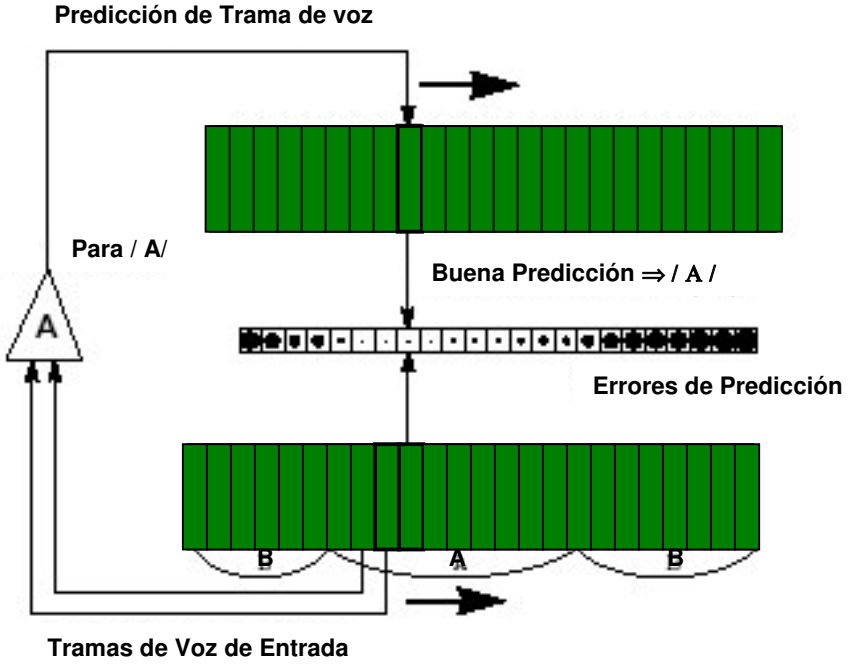


Figura 5.4. Operación Básica de una red Predictiva

La LPNN es un híbrido NN –HMM, lo que significa que el modelamiento acústico es logrado por las redes predictivas, mientras que el modelamiento temporal es logrado por un HMM, lo que implica que la LPNN es un sistema basado en estados donde cada neurona Predictiva corresponde a un estado de un HMM (auto regresivo).

5.2.3 Entrenamiento de la LPNN

La red se entrena sobre un proceso de pronunciación en 3 pasos: un paso forward (en avance), un paso de alineamiento y un paso backward (o de retroceso). Los primeros dos pasos identifican un alineamiento optimo entre los modelos acústicos y la señal de voz (si la pronunciación a sido presegmentada en el nivel de estado estos dos pasos son innecesarios).

Este alineamiento es utilizado para forzar los modelos acústicos durante el paso backward. A continuación se describe el algoritmo de entrenamiento en detalle.

El primer paso es el paso de avance (forward) ilustrado en la figura 6.3 (a). Para cada trama de entrada en tiempo t, se suministra la trama (t – 1) y (t – 2) en paralelo a los reales que están relacionados con esta pronunciación.

Por ejemplo las redes a_1, a_2, a_3, b_1, b_2, b_3 para la pronunciación (aba). Cada red hace una predicción de trama (t) y se calcula la distancia euclideana a la trama actual (t). Estos errores escalares son seleccionados y secuenciados de acuerdo a pronunciaciones conocidas de silabas y almacenados en la columna (t) de una matriz de predicción de error. Esto se repite para cada trama hasta que se halla calculado la matriz completa.

El segundo paso es el de alineamiento, ilustrado en la Figura 5.5 (b). El algoritmo (DTW) estándar es usado para encontrar el alineamiento optimo entre la señal y los modelos de fonemas, como se explicó en capítulos anteriores, identificado por un camino diagonal de avance a través de la matriz de predicción de error, donde el camino más optimo es aquel que tiene la más baja posibilidad de error acumulado.

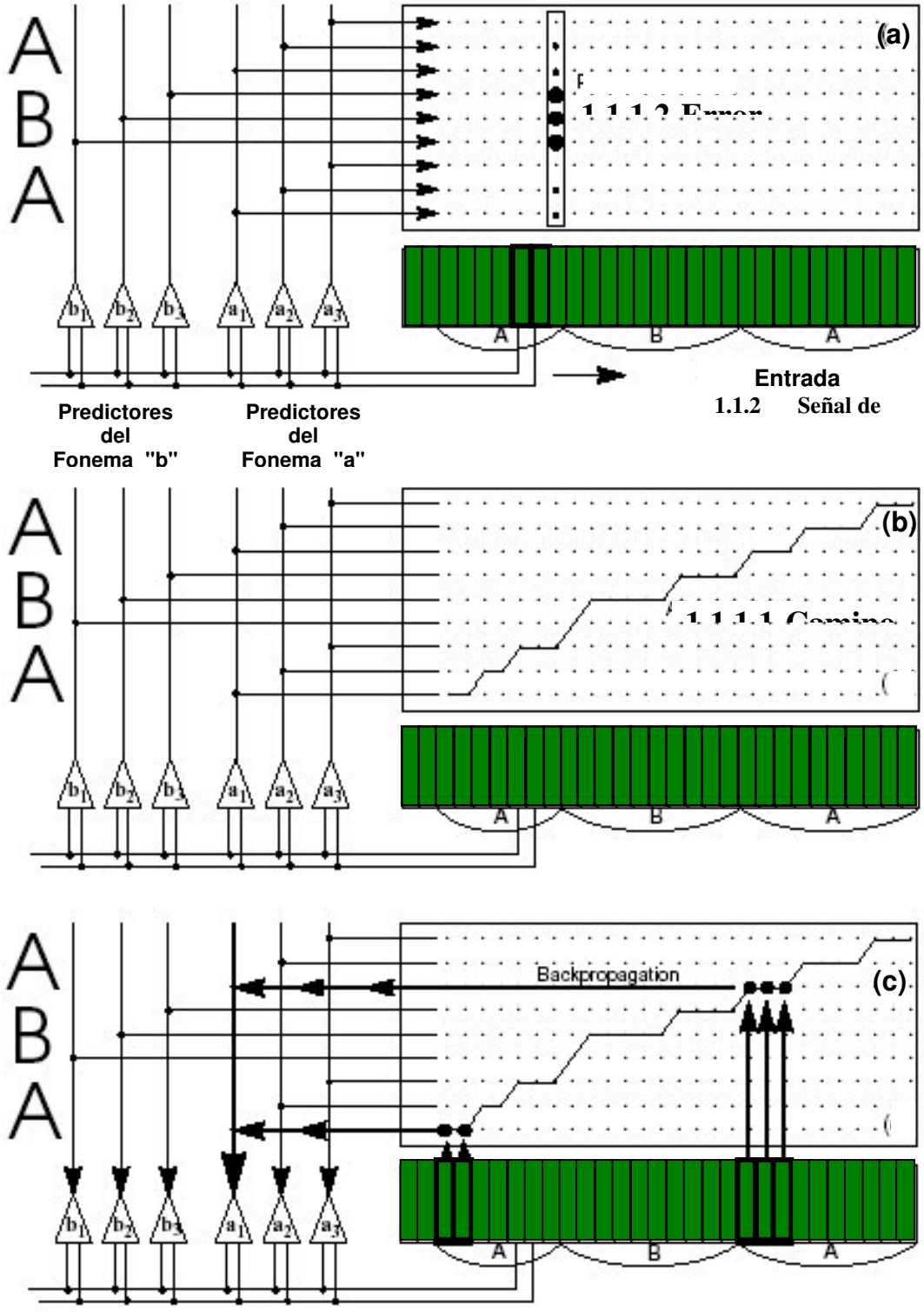


Figura 5.5 Entrenamiento de la red LPNN

El paso final de entrenamiento es el paso backward (o de retroceso) ilustrado en la figura 5.5 (c). En este paso se reprograma el error por cada punto a través del camino de alineamiento, es decir, para cada trama propagamos el error en retroceso para una red, designando la que mejor prediga esta trama de acuerdo al camino de alineamiento; el error retropropagado es simplemente la diferencia entre la predicción de las redes y la trama actual. Una serie de tramas puede retropropogar el error en la misma red como se muestra. El error es acumulado en las redes hasta la ultima trama de la sílaba, tiempo en el cual se actualizan todos los pesos.

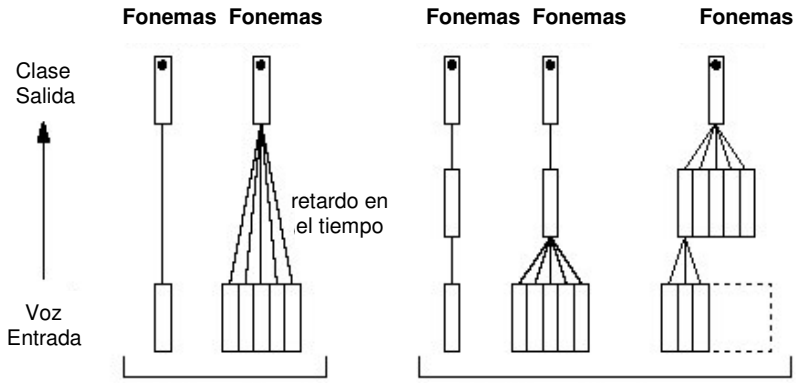
5.3 REDES DE CLASIFICACIÓN

Las redes neuronales pueden entrenarse para mapear un espacio de entrada a cualquier clase de espacio de salida.

Otro tipo útil de *mapping* es la *clasificación,* en la cual los vectores de entrada son mapeados en uno de N clases. Una red neuronal puede representar estas clases por N unidades de salida, de las cuales la uno corresponde a la clase del vector de entrada que tiene un “1” de activación mientras todas las otras salidas un “0” de activación. Un uso típico de esto en el reconocimiento de voz es mapeando las tramas de voz a clases de fonema. La redes de clasificación son atractivas por varias razones:

- Son simples e intuitivos.
- Son naturalmente discriminativos.
- Son modular en el diseño, por lo que pueden combinarse fácilmente en los sistemas más grandes.
- Son matemáticamente entendibles.

- Tienen una interpretación probabilística, por lo que pueden integrarse fácilmente con técnicas estadísticas como HMMs.



1.1.2.1. Perceptron de capa simple

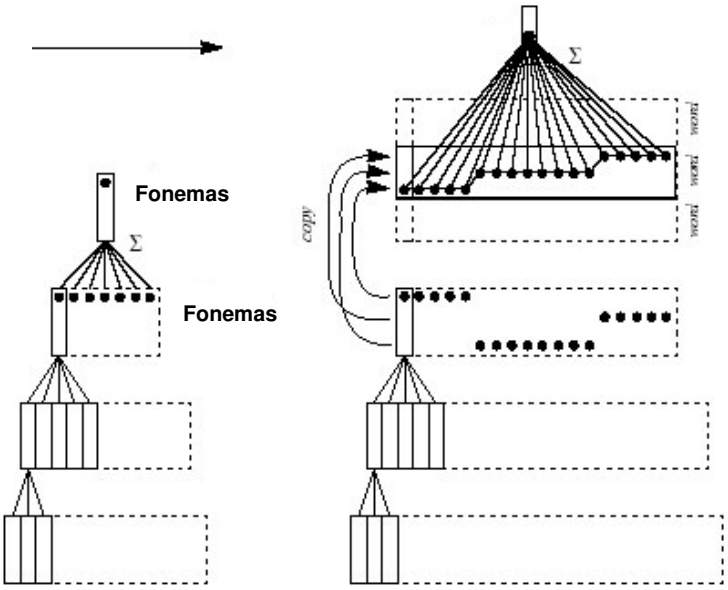


Figura 5.6. Tipos de arquitecturas para redes de clasificación

5.3.1 Arquitectura Perceptron Multicapa MLP

5.3.1.1 Operación Básico

Los Perceptrons constituyen las redes Feedforward mas simples que utilizan aprendizaje supervisado. Una Perceptron consta de unidades *threshold*, dispuestas en capas, como se muestra en la figura 5.7 estas son entrenadas por las reglas delta descrita en secciones anteriores.

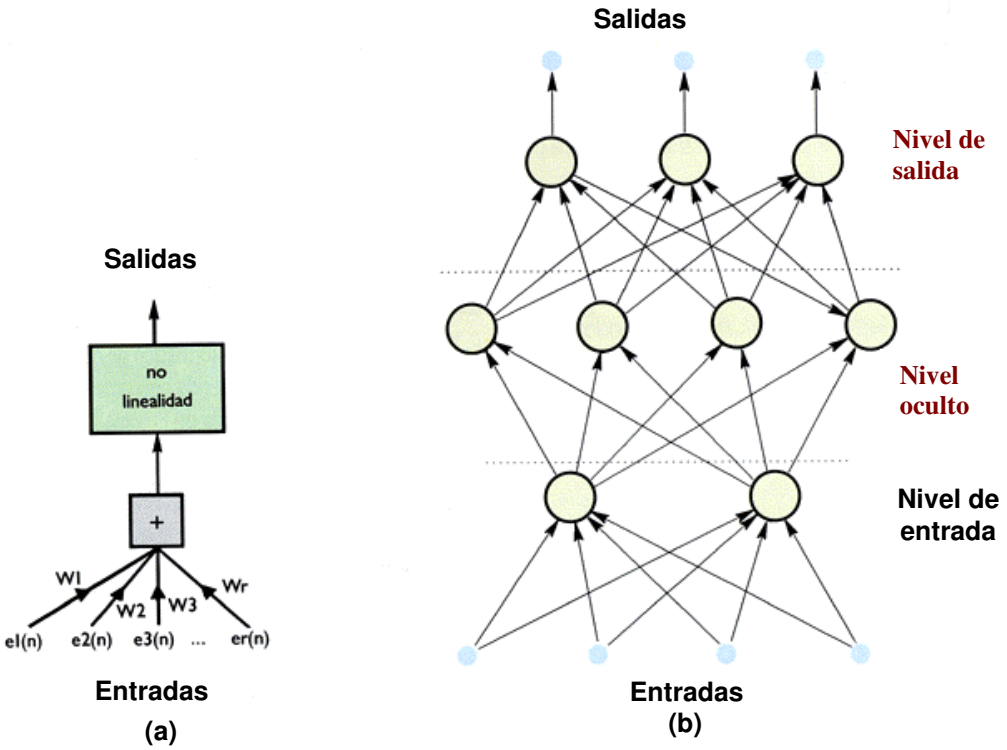


Figura 5.7. Perceptrons. (a) Perceptron simple; (b)Perceptron Multicapa

En el caso de las perceptron simples como se muestra en la figura la regla delta se puede aplicar directamente. Debido a que las activaciones de las perceptron son binarias esta regla de aprendizaje general se reduce a la regla de aprendizaje de la perceptron, que dice que si una entrada es activada (y_i=1), y

la salida y_j es errónea, entonces, w_{ji} debe ser incrementada o decrementada por una pequeña cantidad **ε,** dependiendo si la salida deseada es 1 o 0 respectivamente. Este procedimiento se garantiza para encontrar un peso que clasifique correctamente el patrón en cualquier entrenamiento si los patrones son lineales, separables etc. Sin embargo, la mayoría de los entrenamientos no son linealmente separables, en este caso se requiere de múltiples etapas[11].

Este tipo de redes pueden aprender en teoría, cualquier función, pero son más complejas de entrenar. La regla Delta no se puede aplicar directamente a la MLP porque no hay *targets* en las capas ocultas. Sin embargo, si una MLP utiliza funciones de activación continua en vez de discretas por lo que se hace posible La utilización de derivadas parciales y la regla de la cadena para derivar la influencia de cualquier peso sobre cualquier activación de salida, la cual indica como modificar estos pesos con el fin de reducir el error de la red. la generalización de la regla delta es conocida como *backpropagation*

Las MLP pueden tener varias de capas escondidas, aunque una sola capa escondida es suficiente para muchas aplicaciones y capas ocultas adicionales tienden hacer el entrenamiento mas lento. Las MLP también pueden ser restringidas estructuralmente de varias formas en este caso limitando su conectividad a áreas geométricas locales o limitando el valor de los pesos.

5.3.1.2 Criterios de Entrenamiento

El problema de entrenamiento se especifica como la forma de encontrar la configuración de los parámetros de la red que minimice de alguna forma las funciones de error (que pueden verse como una medida de eficiencia de la red), El error global E en un conjunto particular de datos, en este caso un conjunto de tramas de señal de habla, esta dada por:

[11] BISHOP, Chirstopher M. Neural Networks for Pattern Recognition. 6a Edición. New York: Oxford, 2000. p. 85 – 88.

$$E = \sum_{t=1}^{T} E^{(t)}$$

Generalmente, las bases de estas funciones de error surgen del principio de *Maximum Likelihood* (LM). Para un set de entrenamiento $D = \{u, \varepsilon\}_1^T$ m la probabilidad $\Im$

$$\begin{aligned} \Im &= \prod_{t=1}^{T} p(u^{(t)}, \varepsilon^{(t)}) \\ &= \prod_{t=1}^{T} p(\varepsilon^{(t)} \left| u^{(t)} \right.) p(u^{(t)}) \end{aligned}$$

En vez de maximizar la probabilidad, es mas conveniente minimizar el logaritmo negativo de esta. Donde La función de error se convierte en:

$$\begin{aligned} E &= -Ln\Im, \\ &= -\sum_{t=1}^{-Ln\Im} Lnp(\varepsilon^{(t)} \left| u^{(t)} \right.) - \sum_{t=1}^{-Ln\Im} Lnp(u^{(t)}) \end{aligned}$$

Debido a que el segundo termino no es dependiente de los parámetros de la red, este solo representa una constante que puede separarse de la función de error.

5.3.1.3 Backpropagation

Backpropagation también conocido como Error de retropropagación, o la Regla Delta Generalizada, es el algoritmo de entrenamiento supervisado mas ampliamente utilizado para redes neuronales.

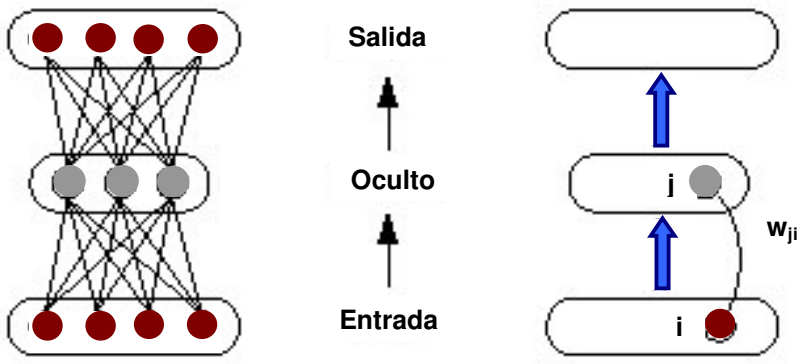


Figura 5.8 Red Neuronal Feedfoward resaltando la conexión de la unidad i a la unidad j

Asumiendo que se tiene una red feedforward de unidades no lineales (típicamente sigmoidales), como la que se muestra en la figura 5.8 se requiere encontrar los valores de los pesos que harán que la red sea capaz de computar una función deseada de vectores de entrada a vectores de salida. Ya que las unidades computan funciones no lineales, no se pueden encontrar estos pesos analíticamente; por lo que se utilizará un procedimiento gradiente descendiente en una función de error global E.

El objetivo de este algoritmo es modificar los pesos de la red neuronal por medio de un proceso iterativo, con el propósito de obtener cada vez valores de salida más similares a los que se desea obtener[12].

El algoritmo *back propagation learning* es un proceso de entrenamiento que consiste de los siguientes pasos:

1. Inicialización de los pesos de la red con valores aleatorios.

[12] Ibid. p. 140.

2. Activación:

- La activación de los nodos de entrada está determinada por los valores obtenidos por la red.
- La activación O_j de los nodos ocultos y los nodos de salida está determinada por:

$$O_j = F(\sum W_{ji} O_i - \theta_j)$$

En donde,

W_{ji} es el peso de una entrada O_i

θ_j es el umbral del nodo y

F es la función:

$$F(a) = \frac{1}{1+e^{-a}}$$

3. Entrenamiento de los pesos:

- Comienza con los nodos de salida y trabaja hacia atrás con los nodos ocultos recursivamente. Los pesos son ajustados mediante:

$$W_{ji}(t+1) = W_{ji}(t) + \Delta W_{ji}$$

En donde:

$W_{ji}(t)$ es el peso del nodo i al nodo j a tiempo t

ΔW_{ji} es el ajuste del peso

- El ajuste del peso es calculado por:

$$\Delta W_{ji} = \eta \delta_j O_i$$

En donde:

$0 < \eta < 1$ y

δ_j es la gradiente de error del nodo *j*

Se puede converger más rápido si se añade el término:

$$W_{ji}(t+1) = W_{ji}(t) + \eta \delta_j O_i + \alpha \left[W_{ji}(t) - W_{ji}(t-1) \right]$$

Donde, $0 < \alpha < 1.$

- El gradiente de error está dado por:

  Para los nodos de salida:

$$\delta_j = O_j (1 - O_j)(T_j - O_j)$$

en donde T_j el valor de activación esperado de salida y O_j es el actual valor de activación de salida del nodo de salida *j*.

Para los nodos ocultos:

$$\delta_j = O_j (1 - O_j) \sum_k \delta_k W_{kj}$$

En donde δ_k es el gradiente de error del nodo *k* al cual apunta una conexión desde el nodo oculto *j*.

Repetir iterativamente hasta que converja al criterio de error seleccionado. Cada iteración incluye los tres pasos anteriores.

5.3.1.4 La MLP como clasificador (estimador de probabilidades)

Fue descubierto recientemente que si un perceptron multicapa es asintoticamente entrenado como un clasificador 1-a-N que usa el error cuadrático medio (MSE, M*ean squared error*) o cualquier criterio similar, por lo que sus activaciones de salida, aproximaran la probabilidad de la clase posterior P(class/input), con una exactitud que mejora con el tamaño del set de entrenamiento. Este hecho importante ha sido probado por Gish (1990), Bourlard & Wellekens (1990), Hampshire & Pearlmutter (1990), Ney (1991), y otros.

Este resultado teórico es empíricamente confirmado en Figura 5.9. Un clasificador de red fue entrenado en un millón de tramas de voz, usando salidas softmax, y se examinaron sus activaciones de salida para ver que tan frecuente cada valor de activación particular era asociado con la clase correcta. Es decir, si la entrada de la red es x, y la k-esima red activación de salida de la red es el $y_K(x)$, dónde el k=c representa la clase correcta, entonces se midió empíricamente $P(k=c|y_K(x))$, o equivalentemente a $P(k=c|x)$, ya que $y_K(x)$ es una función directa de x en la red entrenada. En el gráfico, el eje horizontal muestra las activaciones $y_K(x)$, y el vertical muestra los valores empíricos de $P(k=c|x)$. (El gráfico contiene diez cajas, cada uno con aproximadamente 100,000 puntos de datos.) El hecho que la curva empírica sigue un ángulo de 45 grado indica que las activaciones de la red son de hecho una aproximación cercana para la probabilidad de la clase *posterior*.

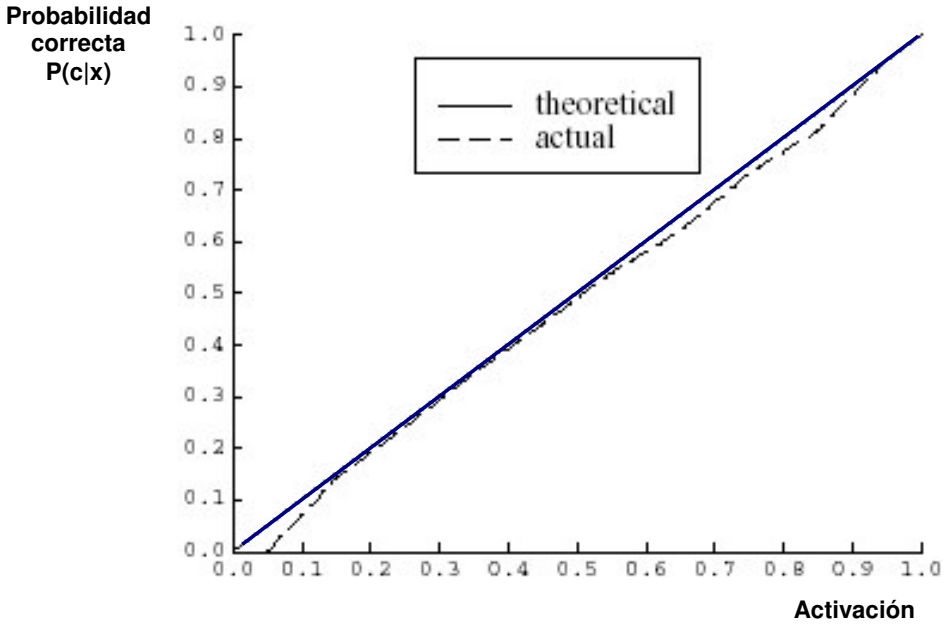


Figura 5.9 Las salidas de activación de la red son estimaciones confiables de las probabilidades de clase *posteriors*

Muchos sistemas de reconocimiento de voz se han basado en DTW aplicados a las clases de activaciones de salida de la red directamente, resaltando las hipótesis mediante la suma de las activaciones a lo largo del mejor camino de alineamiento.

5.3.1.4.1 Las probabilidades Vs. Los *Posteriors.*

La diferencia entre las probabilidades y *posteriors* se ilustran en la Figura 5.10 Suponga tener dos clases, c1 y c2. La probabilidad P(x|ci) describe la distribución de la entrada x dada la clase, mientras el posterior P(ci|x) describe la probabilidad de cada ci dada la entrada. En otros términos, las probabilidades son modelos de densidad independientes, mientras los posterior indican como una distribución de clases dadas se compara con todos los otros. Para las probabilidades nosotros tenemos $\int_X P(x|C_i) = 1$, mientras para los *posterior* se tiene $\sum_i P(c_i|x) = 1$

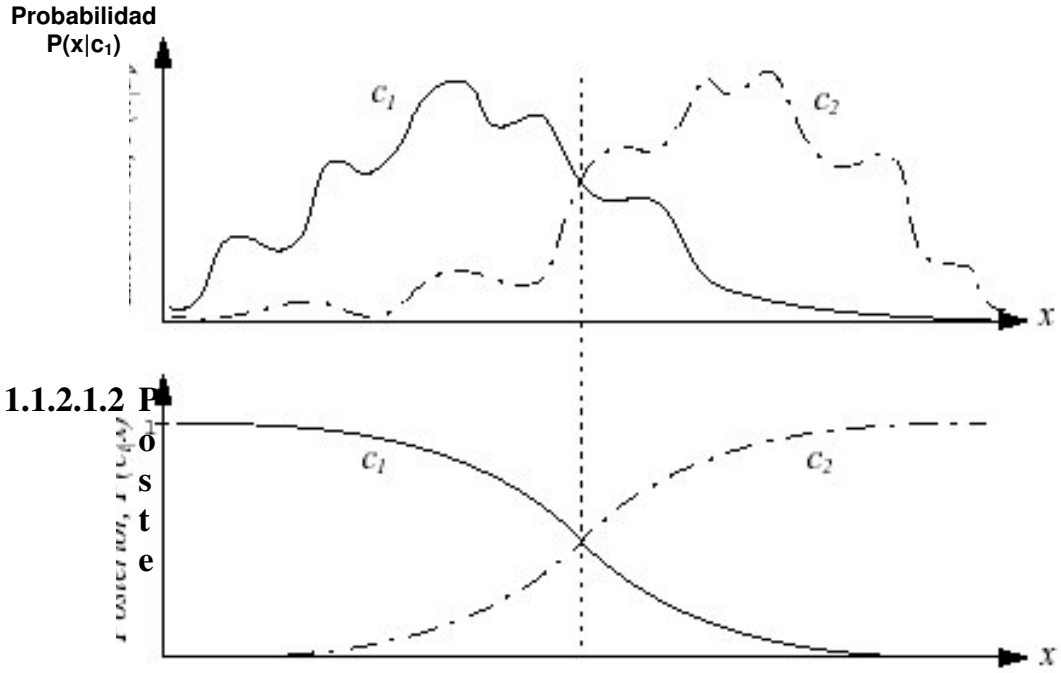


Figura 5.10 Modelo de Densidades independientes; Modelo *Posteriors*

Los *posteriors* son más eficientes clasificando la entrada: la regla de decisión de Bayes dice que se deberia clasificar x en la clase c1 si

$$P(c_1|x) > P(c_2|x)$$

Si se quiere clasificar la entrada usando probabilidades, se tendría que convertir los *posteriors* en probabilidades usando la regla de Bayes, proporcionando una forma más compleja de la regla de decisión Bayes que dice que se debe clasificar x en la clase c_1 si:

$$P(x|c_1)P(c_1) > P(x|c_2)P(c_2)$$

Nótese que los prior $P(c_1)$ están implícitos en los posterior, pero no en las probabilidades, por lo que deben ser explícitamente introducido en la regla de decisión si estamos usando las probabilidades.

Intuitivamente, las probabilidades modelan las superficies de distribuciones, mientras los posterior modelan los límites entre las distribuciones. Por ejemplo, en Figura 5.10, las deformaciones de las distribuciones son modelado por las probabilidades, pero la superficie deformada es ignorada por los posterior, ya que el límite entre las clases está claro sin tener en cuenta las deformaciones. Así, los modelos probabilísticos (como los usados en los estados de un HMM) pueden desperdiciar sus parámetros modelando detalles irrelevantes, mientras los modelos posterior (como los provistos por una red neuronal) puede representar la información crítica más fácilmente.

5.3.2 Redes Neuronales Time Delay (TDNN)

Un tipo de MLP que es bastante relevante para el reconocimiento de habla es la TDNN (*Time Delay Neuronal Networks*). Esta arquitectura fue desarrollada inicialmente para el reconocimiento de fonemas pero también se ha utilizado para el reconocimiento de escritura. La TDNN opera sobre un campo de entrada en dos dimensiones donde la dimensión horizontal es el tiempo y las conexiones son retrasadas en el tiempo. La TDNN tiene tres características especiales:

1. Sus tiempos de retardos son jerárquicamente estructurados, por lo que las unidades de los niveles mas altos pueden integrarse mas en el contexto temporal y mejorar la detección de características en estos niveles más altos.

2. Los pesos están relacionados a lo largo del eje del tiempo, ejemplo: pesos correspondientes a diferentes posiciones temporales comparten los mismos valores, por lo que la red tiene pocos parámetros libres y pueden generalizarse eficientemente.

3. Las unidades de salida integran temporalmente los resultados de los detectores de características locales distribuidos en el tiempo, por lo que la red es invariante, por lo que se pueden reconocer patrones sin importar en que momento ocurran.

La TDNN es entrenada usando el algoritmo de backpropagation estándar lo único que difiere es que la relación de pesos es modificada de acuerdo a la señal de error promedio y no independientemente.

Una TDNN se distingue de una MLP simple no solo por su tiempos de retardo jerárquicos sino también por la integración temporal de activaciones de fonemas sobre varios tiempos de retardo. La TDNN es capaz de clasificar fonemas correctamente aun si estos están segmentados deficientemente.

5.3.3 Red Neuronal *Time Delay* Multi-estado (MS -TDNN)

Una red interesante que combina una red dinámica basada en estados con el fin de aceptar un número variable de tramas de entradas que puede utilizarse para manejar unidades o subunidades de la palabra es la red Neuronal *Time Delay* Multi-estado (Haffner y Waibel, 1992). Como puede verse en Figura 5.11, el MS-TDNN es una extensión del TDNN del nivel fonema al nivel de palabra, y de un solo estado a estados múltiples. Es decir, mientras un TDNN realiza la integración temporal sumando las activaciones de un solo fonema (un solo estado) sobre la duración del fonema, en contraste una MS-TDNN realiza integración temporal aplicando DTW a una secuencia de estados sobre la duración de una palabra.

La figura 5.11(a) muestra un sistema *baseline* que representa una TDNN simple cuyos fonemas de salida son copiados a matriz DTW, la cual mejora el comportamiento de un reconocedor de habla continua. Cabe a anotar que este

sistema es suboptimo porque el criterio de entrenamiento es inconsistente con el criterio de pruebas: la clasificación de fonemas no es clasificación de palabra.

Para manejar esta inconsistencia, mencionada anteriormente, debemos entrenar la red explícitamente para que realice la clasificación de palabra. Con este fin, se debe definir una capa de palabra con una unidad para cada palabra en el vocabulario, como es ilustrado en la Figura 5.11(b) para la palabra particular “gato”. Se correlaciona la activación de la unidad de palabra con el score asociado al DTW estableciendo conexiones del DTW del camino de alineamiento a la unidad de palabra. También, se dan los fonemas con una palabra independiente de pesos entrenables, para reforzar la discriminación de la palabra (por ejemplo, para diferenciar “gato” de “pato” puede ser útil dar el énfasis especial al primer fonema); estos pesos se relacionan sobre todos las tramas en que el fonema ocurre. Así una palabra unidad es una unidad ordinaria, sólo que su conectividad a la capa precedente es determinada dinámicamente, y la entrada de la red debe normalizarse por la duración total de la palabra. La unidad de palabra es entrenada con un *target* de 1 o 0, dependiendo si la palabra es correcta o incorrecta para el segmento actual de voz, y el error resultante es el propagado hacia atrás (*backpropagation*) a través de toda la red. Así, la discriminación de la palabra se trata de manera similar a la discriminación del fonema.

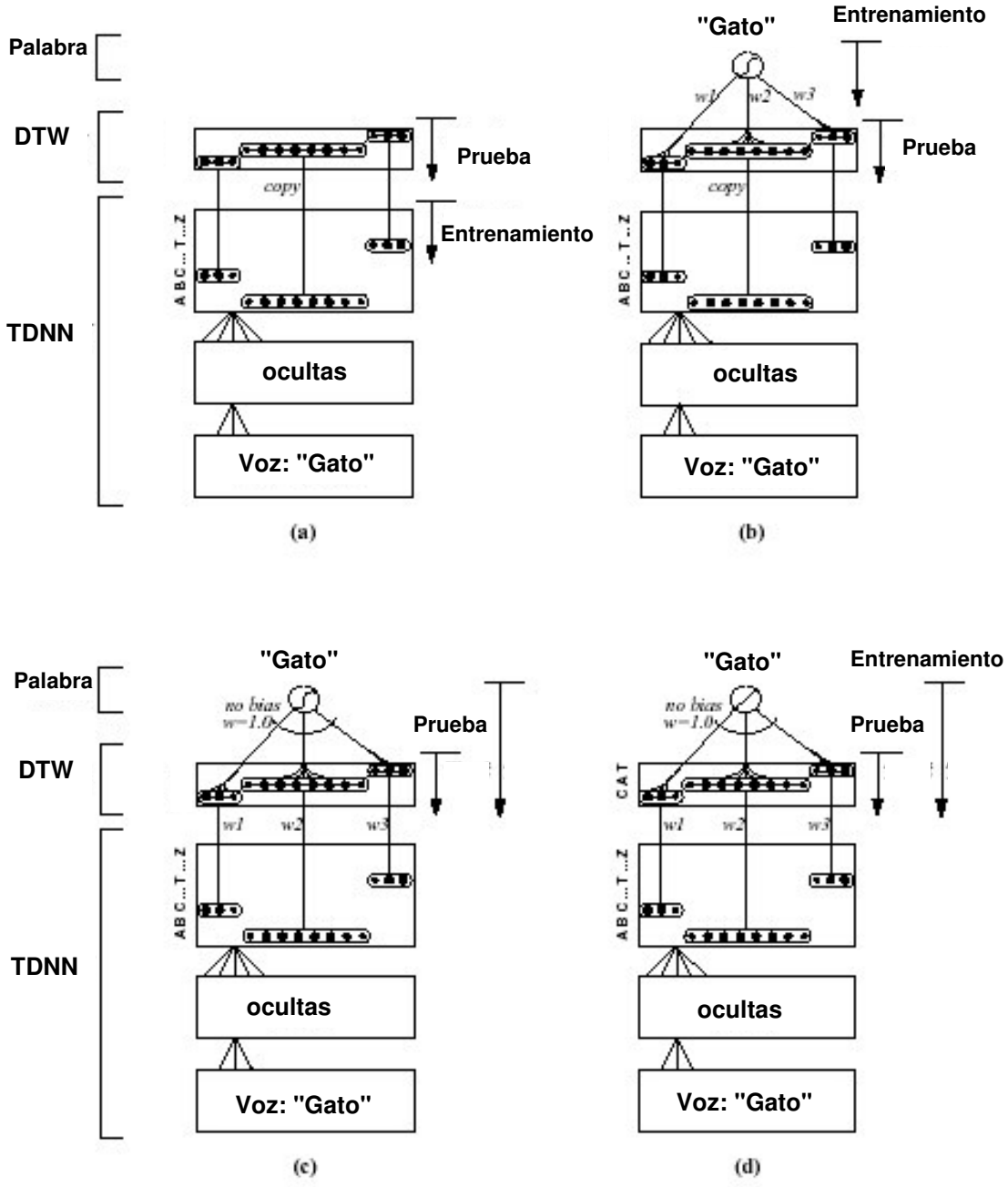


Figura 5.11 MS-TDNN diseñada para resolver inconsistencia

Aunque la red (b) resuelve la inconsistencia original, esta padece ahora una inconsistencia secundaria y es que los pesos guiados a una unidad de palabra son utilizados durante el entrenamiento pero se ignoran durante la prueba, ya

que el DTW aún se está realizando completamente en la capa DTW. Esa inconsistencia se resuelve "bajando" los pesos un nivel, como muestra la Figura 5.11 (c). Ahora las activaciones del fonema no son directamente copiadas en la capa DTW, sino que son modulado por un peso y una ganancia antes de almacenarse (las unidades DTW son lineales); y la unidad de palabra tiene pesos constantes, y no tiene ganancia. Durante el entrenamiento a nivel de palabra, el error es aún retropropagado de los *targets* al nivel de palabra, pero las ganancias y pesos son modificados solamente en el nivel DTW y niveles inferiores. Nótese que esta red transformada, no es exactamente equivalente a la anterior, pero conserva las propiedades de los pesos entrenados asociadas con cada fonema, y hay una ganancia efectiva para cada palabra.

La red (c) todavía se presenta una falla por una inconsistencia menor, que surge de la unidad de palabra sigmoidal. Para reconocimiento de habla continua en el cual se concatenan las palabras en una secuencia, la suma óptima de sigmoides puede no corresponder al sigmoide óptimo de una suma, llevando a una inconsistencia entre la palabra y reconocimiento de la frase. Las unidades de palabra lineales, como el mostrado en la figura 7.28 (d), resolvería este problema; en la práctica se encuentra que las unidades de palabras lineales se comportan ligeramente mejor que las unidades de palabras sigmoidal.

Por lo menos dos inconsistencias potenciales permanecen en la red (d). Primero, el algoritmo de entrenamiento MS-TDNN asume que la conectividad de la red es fija; pero de hecho la conectividad en el nivel de palabra varía, dependiendo de la alineación del camino DTW durante la iteración actual. Como el entrenamiento asintótico y la segmentación se estabiliza, este problema se desprecia.

El MS-TDNN tiene un diseño bastante compacto. Nótese que sus primeras tres capas (el TDNN) es compartido por todas las palabras en el vocabulario, mientras cada palabra requiere sólo un peso y ganancias no compartido para cada uno de sus fonemas. Así el número de parámetros requerido modera

incluso un vocabulario grande, y el sistema puede funcionar con los datos de entrenamiento limitados. Es más, pueden agregarse las nuevas palabras al vocabulario sin volver a entrenar, simplemente definiendo una nueva capa DTW para cada nueva palabra, con los pesos entrantes y las ganancias inicializadas en 1 y 0, respectivamente.

Los pesos constantes dados bajo la capa palabra, pueden argumentar que el entrenamiento a nivel de palabra es simplemente otra manera de ver el nivel de entrenamiento DTW; pero lo anterior es conceptualmente más simple porque hay un solo *target* binario para cada palabra que hace la discriminación en el nivel de palabra muy estrictamente.

6. COMPARACIÓNES ENTRE LAS TÉCNICAS ESTADÍSTICAS Y LAS REDES NEURONALES

Salidas como probabilidades posteriors. Las activaciones de salida de una red de clasificación forma estimaciones muy exactas de las probabilidades *posteriors* P(class|input), de acuerdo con la teoría. Además, estos *posteriors* pueden convertirse en probabilidades P(input|class) para búsqueda de Viterbi más efectiva, simplemente dividiendo la activación por la clase prior P(class), de acuerdo con la regla Bayes[13].

MLP vs TDNN. Una MLP simple mejora la exactitud de palabra en comparación con un TDNN con las mismas entradas y salidas, donde cada uno está entrenado como un clasificador de trama que usa una base de datos grande. Como los tiempos de retardo son redistribuidos en capas superiores dentro de una red, cada unidad oculta ve menos contexto, por lo que se vuelve más simple y menos potente en el reconocimiento de patrones; sin embargo esta también recibe más entrenamiento debido a que este se aplica sobre varias posiciones adyacentes (con pesos relacionados) para que esta aprenda los patrones más simples con más fiabilidad. Así, cuando los datos de entrenamiento son relativamente pequeños como en las pruebas en el reconocimientos de fonemas (Lang 1989, Waibel et en 1989) el tiempo de retrasos jerárquico sirve para aumentar la cantidad de datos entrenados por pesos y mejorar la exactitud del sistema. Por otro lado, cuando se dispone de una gran cantidad de datos de entrenamiento el tiempo de retraso jerárquico de un TDNN hace la unidades ocultas innecesarias y degrada la exactitud del sistema, por lo que un simple MLP es preferible.

[13] Ibid. p. 17 - 23

HMM Vs. NN

Teniendo en cuenta las ventajas y desventajas de cada una de las técnicas expuestas en el desarrollo de esta monografía se pueden hacer las siguientes comparaciones:

Exactitud en el Modelamiento La Densidad discreta de los HMMs padecen errores del cuantización en su espacio de entrada, mientras densidad continua o semi-continua de los HMMs sufren un no emparejamiento de modelos, por ejemplo, un emparejamiento poco eficiente entre un priori de un modelo estadístico (por ejemplo, una mezcla de K Gaussianas) y la verdadera densidad de espacio acústico. En contraste, las redes neuronales son modelos no paramétricos que ni padecen de error de cuantización ni hace suposiciones detalladas sobre la forma de la distribución para ser modelada. Así una red neuronal puede formar modelos acústicos más exactos que un HMM.

Sensitividad de Contexto. Los HMMs asumen que las tramas de voz son independientes de cada otra, para que ellos puedan examinar sólo una trama a la vez. Para tomar ventaja de la información contextual en las tramas vecinas, los HMMs debe artificialmente absorber aquellas tramas en la trama. En contraste, las redes neuronales pueden acomodar cualquier tamaño de la ventana de entrada, porque el número de pesos requeridos en una red simplemente se incrementa linealmente con el número de entradas. Así una red neuronal puede discriminar mas naturalmente el contexto sensitivo que un HMM.

Discriminación. El criterio HMM de entrenamiento estándar (LM), no diferencia explícitamente entre los modelos acústicos. Es posible mejorar la discriminación en un HMM usando el criterio de Máxima Información Mutua, pero este es más complejo y difícil de implementar. En contraste, la discriminación es una propiedad natural de las redes neuronales cuando son entrenadas para

realizarla clasificación. Así una red neuronal puede diferenciar más naturalmente que un HMM.

Economía. Un HMM utiliza sus parámetros para modelar la superficie de la función de densidad en el espacio acústico, en términos de las probabilidades P(input|class). En contraste, una red neuronal usa sus parámetros para modelar los límites entre las clases acústicas, en términos de los posterior P(class|input). Cada superficie o límite puede utilizarse para clasificar la voz, pero los límites requieren menos parámetros y así puede hacer mejor uso de datos de entrenamiento limitados. Así una red neuronal es menos costosa que un HMM.

Los HMMs se conocen por su impedimento debido a la suposición de primer orden, por ejemplo, la suposición que todo las probabilidades dependen del estado actual, independiente de los datos anteriores; limita la habilidad del HMM de modelar los efectos de coarticulación, o para modelar las duraciones con precisión. Desafortunadamente, los híbridos NN-HMM comparten este impedimento, porque la suposición de Primer orden es una propiedad del modelo temporal HMM, no del modelo acústico de la red neuronal

7. APLICACIONES DE LA TECNOLOGIA CSR

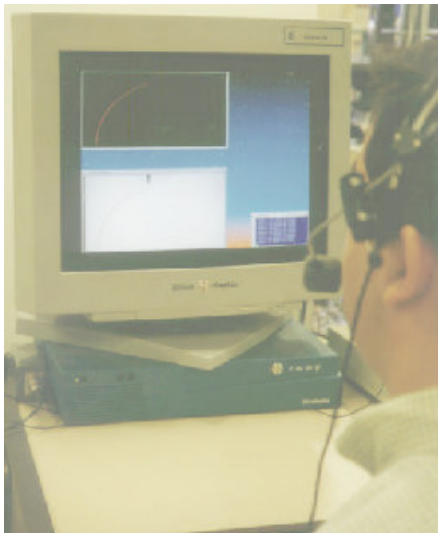
A la hora de desarrollar aplicaciones con sistemas de CSR hay que tener en mente que con la tecnología actual los sistemas no están exentos de errores, por lo que las primeras aplicaciones en las cuales este tipo de interface comienza a tener éxito son aquellas que se caracterizan por ser simples, en cuanto es sencillo el uso del mismo, supone una evolución de la tecnología ya existente en el sentido de que únicamente realizamos un cambio de interface y sobre todo, la aplicación debe ser tolerante a errores. A estas consideraciones sobre la aplicación en sí, hay que añadir una serie de requerimientos tecnológicos del sistema de CSR.

Para trabajar en aplicaciones reales, el sistema de CSR tiene que tener la capacidad de reconocer palabras o comandos de la aplicación en un contexto de habla fluida, mantener un nivel de prestaciones adecuado frente a cambios de usuario, canal de comunicación, ruidos, etc., permitir el rechazo de palabras que no formen el vocabulario de reconocimiento y trabajar en tiempo real entre otros requerimientos.

Actualmente hay un gran número de aplicaciones potenciales que pueden usar el Reconocimiento de habla continua. Algunas de ellas son viables en la actualidad y otras tendrán que esperar mejoras en la potencia de los reconocedores o reducciones de coste.

Algunas aplicaciones requieren solo un pequeño vocabulario de palabras aisladas pronunciadas por un único locutor, mientras otras requieren comprensión del mensaje pronunciado por cualquier locutor. Algunas requieren

una respuesta inmediata del sistema, mientras que en otras se puede tolerar un pequeño retraso. En casi todas las aplicaciones una tasa de reconocimiento del 100% es deseable, pero en algunas se puede permitir un pequeño porcentaje de error.



Como ejemplo puede citarse el sistema AUDIOTEX, desarrollado en los laboratorios de Telefónica I+D para ser implantado en la red telefónica conmutada, y que ofrece un servicio de noticias en el que el usuario selecciona con su voz la información que desee. Este sistema integra un reconocedor de palabras aisladas independiente del locutor, basado en HMMs continuos. Durante el diseño y las pruebas de campo de este sistema, los autores han podido comprobar la importancia de la gestión de diálogos y, en general, del tema de los factores humanos, en el éxito y aceptación por el gran publico de este tipo de sistemas; se ha visto que aunque la tasa de reconocimiento en laboratorio supera fácilmente el umbral del 96%, las pruebas reales del sistema con usuarios no entrenados, pueden generar una tasa de fallos muy superior a lo esperado cuando la gestión del dialogo que incita al usuario a pronunciar las palabras que han de ser reconocidas no se estudia a fondo. Tienen pues, mucho que decir en este tema los psicólogos y los estudiosos del comportamiento de la mente humana en general, ya que al final, es de ellos de quien depende que este tipo de tecnologías salga algún día de los laboratorios para implantarse en la vida real.

7.1 TIPOS DE APLICACIONES

Se pueden dividir las aplicaciones en tres tipos:

- Entrada de datos. Por ejemplo, datos en una inspección de fabrica, o el rellenar un informe medico.

- Control de sistemas. El control de una centralita telefónica por medio de la voz, o el panel de mandos de un avión, el control de acceso como se demostró en este trabajo son ejemplos de estas aplicaciones.

- Comunicaciones, que engloba desde el encaminamiento del mensaje hasta el contenido del mensaje mismo. Se podría pensar en una marcación por voz como la más elemental, o la sustitución de una operadora como la más compleja.

7.1.1 Áreas de Aplicación

Son distintas las áreas en que se puede aplicar esta tecnología:

Área profesional. En este tipo de aplicaciones, los requerimientos del sistema de reconocimiento pueden ser menores, ya que se puede entrenar para las personas que lo usen. Los ambientes pueden variar desde el más silencioso de las oficinas, a los más ruidosos de las fabricas. Normalmente las personas que utilizan estos sistemas lo usaran con cierta frecuencia y serán expertos en su uso. Aplicaciones típicas son el manejo de un ordenador por medio de la voz, o incluso el control de una máquina herramienta por medio de la voz.

Área publica. Se entiende por esto el tipo de aplicaciones en que el sistema puede ser usado por cualquier persona, lo que requerirá que el reconocedor de habla sea independiente del locutor. Normalmente se piensa en aplicaciones que den servicio a través de la red telefónica, lo que añade problemas al

reconocedor, debido fundamentalmente al recorte de banda y a la gran cantidad de entornos desde donde se puede hacer una llamada telefónica. Son muchas las aplicaciones que se pueden agrupar en esta área, desde las más elementales de manejo de un sistema de correo de voz por medio de comandos orales, hasta las más complicadas de sustitución de una operadora, pasando por accesos a bases de datos remotas.

Una de las área con más aplicaciones potenciales son las telecomunicaciones y servicios añadidos. En ciertos servicios añadidos a la red telefónica, el uso de interfaces orales permite una reducción efectiva del coste del servicio. Ejemplos de estas aplicaciones son la automatización de los servicios de operadora y la validación de compras con tarjetas de crédito. En el primer caso, existen aplicaciones en uso en los EE.UU. por parte de las compañías telefónicas AT&T y Northen Telecom para automatizar el servicio de facturación de llamadas asistidas por operadora. En estos casos, el reconocimiento del mensaje se realiza mediante un sistema de localización de palabras. En el caso de validación de compras con tarjeta de crédito, este servicio es utilizado por comercios que no disponen de módems para validar la venta. Con un sistema de reconocimiento de dígitos conectados puede reconocerse los números de la tarjeta de crédito, la identificación del vendedor y el valor de la venta. Como el número de la tarjeta de crédito y la identificación del vendedor están formados por una secuencia de dígitos con ciertas restricciones, no causan problemas a la hora de reconocerlos. La incorporación de interfaces orales ha permitido también incrementar el número de servicios proporcionados por una red de telecomunicaciones. Ejemplos de estas aplicaciones son los servicios de información y transacciones bancarias, servicios de telefonía interactiva (p.e. el sistema VIP -Voice Interactive Phone- de AT&T que permite acceder a ciertos servicios pronunciando el nombre asignado al mismo en lugar de pulsar un código con el teclado multifrecuencia) y servicios de acceso a información (p.e. sistemas de audiotex).En relación a la telefonía móvil en vehículos, los sistemas

de reconocimiento de voz comienzan a ser introducidos para permitir controlar el teléfono (funciones de marcado, respuesta, etc.) mediante comandos orales.

Aplicaciones militares. Donde existen posibilidades de manejo de barcos, aviones o armas por medio de la voz. Hay un área potencial de utilización de estas técnicas en el campo de la inteligencia militar para localizar mensajes o palabras claves.

En aplicaciones militares se ha experimentado en la introducción de interfaces orales para interactuar con los sistemas básicos de un avión de guerra. Los sistemas de reconocimiento suelen ser capaces de dar unas prestaciones muy buenas trabajando con relaciones señal a ruido muy pequeñas. En la aviación civil también se pueden encontrar aplicaciones en proceso de experimentación para el control aéreo utilizando sistemas de reconocimiento de habla continua.

Ayudas a discapacitados. Una de las aplicaciones más inmediatas de los sistemas de CSRs como interfaz entre hombre y máquina es la ayuda a discapacitados físicos. Mediante comandos orales se pueden controlar muchas de las funciones y actividades cotidianas. Ejemplos en fase de experimentación son la silla de ruedas controlada oralmente, camas hospitalarias, control oral de teléfonos (ej. listas telefónicas controlado oralmente) y la activación oral de aparatos y sistemas domésticos. En el caso del teléfono controlado oralmente, el usuario puede almacenar y acceder a una lista de números telefónicos utilizando comandos orales. En este tipo de aplicaciones, el sistema de reconocimiento de voz es dependiente del locutor y trabaja normalmente bajo la configuración de reconocimiento de palabras conectadas con capacidad de localización de los comandos en habla extraña . El sistema tiene que tener la capacidad de ser entrenado por el usuario para de esta forma hacer el acceso a los números telefónicos mediante el nombre de la persona que se desea llamar.

La activación oral de aparatos y sistemas domésticos, incluida dentro del campo de la domótica, tiene como objetivo el controlar a estos mediante comandos

orales a través de un sistema de diálogo. Son susceptibles de control oral, aparatos como el televisor (encender/apagar, cambiar de canal, volumen), el equipo de HIFI, abrir y cerrar puertas, abrir y cerrar persianas, control de una cámara de seguridad, activar el teléfono, la calefacción, el horno y encimera, encender y apagar luces, etc. En 1984, la empresa británica Voice Input Systems construyó, demostró y comenzó a comercializar el sistema VADAS para ayudar a discapacitados físicos a controlar oralmente dispositivos domésticos. Una capacidad interesante de estos sistemas de control oral de dispositivos domésticos es la posibilidad de controlarlos de forma remota a través de la línea telefónica. Los sistemas de reconocimiento utilizados en este tipo de aplicaciones suelen ser de palabras aisladas con la capacidad de rechazar habla o sonidos extraños y dependientes del locutor, de modo que se entrena el sistema con la voz del usuario.

APENDICE A.

BASES DE DATOS

Los sistemas de reconocimiento de habla emplean patrones de palabras o de unidades inferiores (como por ejemplo, fonemas) de forma que por medio de algoritmos de comparación de los patrones con las señales acústicas de entrada, se identifican las palabras pronunciadas por un locutor.

Los patrones empleados deben ser suficientemente representativos. Esto es, deben modelar de la forma más exacta posible las características particulares del objeto que representan (palabras, fonemas, etc.), así como todas las posibles fuentes de variabilidad (acento, sexo, edad, dialecto, duración de los sonidos, tipo de micrófono, o características de la línea telefónica utilizada), pues sólo de esta forma el sistema se comportará de forma correcta en las situaciones reales en las que debe funcionar.

Para poder obtener patrones que cumplan estos requisitos, es necesario disponer de bases de datos de voz que contemplen todas las fuentes de variabilidad.

Una base de datos de voz consiste en un conjunto de ficheros que contienen cada uno alguna palabra o frase pronunciada por un locutor, junto con la transcripción exacta de su contenido y otras informaciones adicionales como la relación señal-ruido (SNR), el sexo del locutor, su edad (si es un niño o un adulto) y el tipo de ruido que hay presente en el fichero (ruido telefónico, voces de fondo, ruido ambiental o golpes repetitivos).

El proceso de obtención de una base de datos de voz se puede dividir en dos partes:

1. Captura de la base de datos de voz, que es la fase en la que se graban palabras y/o frases pronunciadas por un número suficientemente grande de personas.

2. Etiquetado de la base de datos. En esta fase se añadirá a cada grabación, la transcripción exacta de su contenido e información adicional que pueda ser útil cuando se vaya a trabajar con la misma ver Figura 6.1.

Las características y composición de la base de datos de voz depende del reconocedor que se desee diseñar. Así pues, para desarrollar un reconocedor que reconozca números del cero al nueve se necesitará una base de datos que esté compuesta por ficheros de voz que tengan pronunciaciones de dígitos del cero al nueve, y para diseñar un reconocedor de vocabulario grande y configurable, la base de datos requerida deberá contener gran variedad de palabras distintas, de forma que se puedan modelar patrones de voz para reconocer cualquier palabra del castellano.

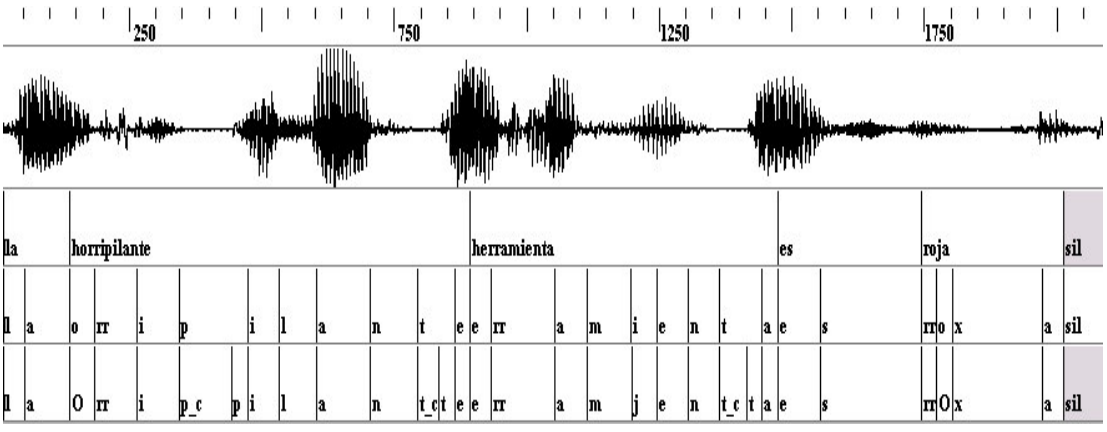


Figura 6.1 Transcripción Fonética

El proceso de diseño de un sistema de reconocimiento conlleva un elevado grado de dificultad. Como se observa al describir los diferentes módulos que intervienen en un sistema de reconocimiento, lo más característico del sistema es la fase de reconocimiento de patrones. Por tanto, lo más característico del proceso de diseño será la definición y obtención de los modelos o patrones de unidades lingüísticas.

Los pasos necesarios para llegar a los patrones de unidades lingüísticas comienzan con la obtención de un conjunto de datos de entrenamiento, es decir, una base de datos de voz compuesta por varias pronunciaciones de las unidades a modelar. Junto a esa base de datos es preciso disponer de otra que permita evaluar el comportamiento del sistema trabajando con los modelos entrenados. Ambas bases de datos, de entrenamiento y de evaluación, deben reunir importantes propiedades si se diseñan sistemas independientes del locutor:

La base de datos de entrenamiento debe ser lo suficientemente amplia para contener un número de repeticiones de cada unidad, en diferentes contextos (para suficiente número de hablantes y canales de transmisión distintos), que permita una estima estadísticamente consistente de los parámetros que intervienen en el modelo de dicha unidad.

La base de datos de evaluación debe contener el suficiente número de repeticiones de cada palabra del vocabulario, en diferentes contextos (para suficiente número de hablantes y canales de transmisión distintos), como para que la estima de las tasas de error del sistema sean estadísticamente fiables.

Una vez diseñadas las bases de datos de entrenamiento y evaluación, el paso siguiente irá encaminado a la obtención de la mejor topología y valores de los parámetros correspondientes para los modelos de unidades. Un largo proceso,

que habitualmente se realiza recurriendo a una base de datos de prueba obtenida como un subconjunto de la de evaluación.

Evidentemente, para aquellos sistemas basados en modelos de palabra, cada vez que se desee reconocer un nuevo vocabulario se necesitará grabar una nueva base de datos y repetir el proceso de entrenamiento. Sin embargo en un reconocedor basado en unidades inferiores a la palabra puede pensarse en disponer de un procedimiento de entrenamiento que permita al sistema un funcionamiento independiente del vocabulario; es decir, sin necesidad de repetir totalmente el proceso de entrenamiento cada vez que se cambie el vocabulario de palabras a reconocer.

Algunas bases de datos existentes en la actualidad que tienen como objetivo brindar un soporto al desarrollo de reconocedores de voz en español son:

VESTEL. (Voz en español por vía telefónica), esta es una base de datos de voz recogida en la División de Tecnología de Habla de Telefónica Investigación y Desarrollo de para la implementación de reconocedores de voz en español, independientes de locutor, basados tanto en el reconocimiento de palabras como en unidades inferiores a la palabra. Más de 16.000 personas llamaron a la aplicación respondiendo a una campaña publicitaria lanzada en periódicos locales de toda España. Al llamar a la aplicación, una voz pregrabada les pedía que pronunciaran dígitos del cero al nueve, números, comandos, la provincia desde la que llamaban, su provincia de nacimiento, su nombre y apellidos, y les pedía también que deletrearan algunas palabras.

APENDICE B

SOFTWARE DE RECONOCIMIENTOS DE HABLA CONTINUA

PROGRAMA 1. Speech Servers

- Aplicaciones de IVR con Reconocimiento Natural del Habla

Estas aplicaciones permiten a los usuarios de los sistemas de IVR obtener información y realizar transacciones simplemente hablando en forma natural. Anteriormente sólo se podía interactuar con los IVR por medio de discado (tonos DTMF y/o pulsos decádicos) o bien por aplicaciones de reconocimiento de voz que sólo brindaban la posibilidad de reconocimiento discreto de números y reconocían un rango acotado de palabras en forma discontinua, es decir, sólo de a una palabra a la vez.

- SpeechWoks

La nueva tecnología de *SpeechWorks* permite el reconocimiento natural y continuo del habla sin necesidad de pausa o enunciados artificiales entre cada palabra. Esta nueva tecnología funciona en forma independiente de quien habla, por lo tanto no es necesario ningún tipo de entrenamiento previo. *SpeechWorks* se basa en el reconocimiento de fonemas, lo cual permite incrementar la exactitud de comprensión mediante el manejo de extensos vocabularios. También utiliza técnicas de modelización que otorgan mayor precisión en base a la comparación de sonidos reconocidos contra una lista de "reglas gramaticales", que determinan la correcta interpretación de las palabras, lo cual permite reconocer enunciados o frases completas .

El reconocimiento natural del habla e interfases de usuario avanzadas, permiten

la conducción de “diálogos interactivos” con los usuarios para completar transacciones que potencializan el diseño de las más robustas y versátiles aplicaciones jamás desarrolladas en la industria de los IVR´s. *SpeechWorks* brinda una tecnología vanguardista con un vocabulario de más de 70.000 palabras. Las órdenes sencillas son reconocidas en más de un 98%. Aún los vocabularios más amplios y complejos se manejan con una exactitud mayor al 95% en casos reales.

Actualmente *SpeechWorks* soporta múltiples lenguajes, los más utilizados son: Inglés, Español, Portugués, Francés y Alemán. Adaptándose también a los diferentes acentos regionales.

Software disponible para investigación

- HTK. Universidad de Cambridge

Programa 1. **Hidden Markov Model Toolkit**

http://htk.eng.cam.ac.uk/download.shtml

Aquí pueden encontrarse los fuentes en C del conjunto de herramientas clásico para la construcción de reconocedores automáticos de habla basados en modelos ocultos markovianos.

- **Festival. Universidad de Edimburgo.**

1.1.1.1.1 Programa 2. ***Text-to-Speech***

http://www.cstr.ed.ac.uk/projects/festival.html

Software de distribución libre construido sobre la base de las "speech-tools" de la Universidad de Edimburgo y con contribuciones de diversas instituciones académicas.

- **Festvox. Universidad de Carnegie Mellon.**

Programa 3. **TTS Festival**

http://www.festvox.org/index.html

En este sitio pueden encontrarse un conjunto de herramientas para agregar nuevas voces al TTS Festival.

- **CSLU. Center of Speech and Language Understanding. Oregon.**

Programa 4. http://cslu.cse.ogi.edu/toolkit/

Ofrece un conjunto de herramientas para desarrollar sistemas de diálogo. El toolkit está constituido por una versión del Festival, herramientas para la construcción de aplicaciones con modelos markovianos ocultos y redes neuronales entre otras.

- **EMU. Universidad de Sidney.**

Programa 5. The EMU Speech Database System

http://www.shlrc.mq.edu.au/emu

Es un conjunto de herramientas para analizar, etiquetar la señal acústica y almacenar los parámetros en una base de datos.

- **Wavesurfer.**

Programa 6. **Wavesurfer**

http://www.speech.kth.se/wavesurfer/

El programa Wavesurfer es una herramienta útil para analizar y etiquetar las formas de onda en varios formatos. Permite visualizar el espectro, el espectrograma y la frecuencia fundamental.

CONCLUSIONES

Esta monografía se encamino a responder la pregunta de que si la red neuronal puede servir como un fundamento útil para un sistema de reconocimiento de grandes vocabularios, habla continua y reconocimiento de locutores.

La importancia del

Las Redes neuronales como Modelos Acústicos

Un sistema de reconocimiento de voz requiere las soluciones a los problemas del modelamiento acústico y temporal. La tecnología de reconocimiento de voz ha prevalecido, los modelos ocultos de Markov, ofrece soluciones a los dos problemas: el modelamiento acústico se proporciona por modelos de densidad discreto, continuo, o semicontinuo y el modelamiento temporal se proporciona por estados conectados por transiciones, colocados en una jerarquía estricta de fonemas, palabras, y frases.

Aunque las soluciones de HMM son eficaces, padecen varios inconvenientes. Específicamente, los modelos acústicos padecen errores del cuantización y/o una adopción de modelamientos de parámetros ineficentes; el criterio de probabilidad Máxima estandar (Maximum Likelihood) conduce a una vaga discriminación entre los modelos acústicos; la suposición de independencia se convierte en un obstáculo para el aprovechamiento de las múltiples tramas de

entradas. Dado que los HMMs presentan estos inconvenientes se hizo necesaria la consideración de soluciones alternativas.

Las redes neuronales conocidas por su habilidad de aprender las funciones complejas, generaliza eficazmente, tolerar el ruido, y soportar el paralelismo, ofrece una alternativa prometedora. En un sistema de reconocimiento de voz, se suele usar las redes neuronales para el modelamiento acústico, pero no para modelamiento temporal. Basado en estas consideraciones, se han expuestos sistemas hibridos NN-HMM, en los cuales las redes neuronales son responsables del modelamiento acústico, y HMMs se encargan del modelamiento temporal.

Se expusieron dos formas de usar las redes neuronales para modelamiento acústico. El primero era una nueva técnica basada en la predicción (Enlazamiento predictivo de redes neuronales, o LPNN), en el cual cada clase del fonema fue modelado por una red neuronal separada, y cada red intenta predecir la próxima trama de voz dados algunos tramas recientes de voz; el error de predicción se utilizó para realizar una búsqueda Viterbi para la mejor secuencia, como en un HMM. Se encontró que esta técnica padeció una falta de discriminación entre las clases de fonema, como todas las redes en aprendizaje para realizar un mapeo cuasi-idéntico entre las trama cuasi-estacionaria de su respectivas clases de fonemas.

La segunda técnica está basada en la clasificación en donde una sola red neuronal intenta clasificar un segmento de voz en su clase correcta. Este técnica ha demostrado ser mucho más exitoso, ya que se basa en la discriminación entre las clases de fonema.

Cabe anotar que la mayor parte del material recopilado para esta monografía fue documentación en ingles, debido a que por ser este un campo de investigación relativamente nuevo, la información mas confiable en esta área,

tanto la experimentación como la teoría se encuentra en este idioma, por lo que este trabajo se proyecta a futuras investigaciones que involucren el reconocimiento de habla continua.

BIBLIOGRAFIA

- A.B. Poritz, "*Hidden Markov Models: A Guided Tour*", ICASSP'88, Nueva York, Estados Unidos: 1988, p. 7-13.

- BISHOP, Chirstopher M. Neural Networks for Pattern Recognition. 6a Edición. New York: Oxford, 2000. p. 85 – 88.

- FREEMAN, James A, SKAPURA, David M. Redes Neuronales algoritmos, aplicaciones y técnicas de propagación. Editorial Addison – Wesley Iberoamérica, S.A. 1993.

- HILERA, José R. y MARTÍNEZ, Victor J. Redes Neuronales Artificiales. Madrid, España: Alfa-Omega Editor, 2000. p. 51 – 56

- LINGGARD, MYERS AND NIGHTINGALE. Neural Networks for vision, speech and natural language. Editorial Chapman y Hall, 1992. Capitulo 2, Speech, pagina 129.

- L. R. Rabiner y S. E. Levinson, "Isolated and Connected Word Recognition- Theory and Selected Applications", IEEE Tratados sobre Comunicaciones, Vol. COM-29, Nº . 5, 1981, pp. 121-129.

- PICONE, J.W. , "Signal modeling techniques in speech recognition," *IEEE*, 1993. p. 4 .

- VELÁSQUEZ SILVA, Juan Domingo. "Análisis fino del tracto vocal basado en filtros LPC aplicado al mejoramiento de la calidad de síntesis de voz". Universidad de Chile, Departamento de Ciencias de la Computación. 1996.

- VIDAL, José E. Decodificación Acústico-Fonética mediante plantillas Subléxicas. España: Comisión Internacional de Ciencia y Tecnología, 1990 p. 1-2

- www.pue.udlap.mx/~tesis/lis/lopez_m_j.html

- http://www.is.cs.cmu.edu/%7emonika/thesis/html/Thesis E.html

- http://www.imim.es/quark/num21/021063.htm

- http://www.imim.es/quark/Num21/021062.htm

- http://www.isip.msstate.edu/publications/courses/ece_8463/lectures/current/

- http://liceu.uab.es/~joaquim/speech_technology/tecnol_parla/recognition/refs_reconeixement.html

- http://cslu.cse.ogi.edu/HLTsurvey/ch1node3.html