UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE INGENIERIA
DIVISION DE ESTUDIOS DE POSGRADO

SISTEMAS DE RECONOCIMIENTO DE PALABRAS AISLADAS Y CONECTADAS USANDO LA TRANSFORMADA DE KARHUNEN-LOEVE

T E S I S

QUE PARA OBTENER EL GRADO DE:

DOCTOR EN INGENIERIA

(E L E C T R I C A)

P R E S E N T A :

JOSE ABEL HERRERA CAMACHO

DIRECTOR DE TESIS: DR. RALPH ALGAZI

CIUDAD UNIVERSITARIA, MARZO DEL 2001

A mi esposa Nora y a mis hijos Erika y David, por ser la parte más importante de mi vida.

A mis padres Felisa y Abel por nunca dejar de ser los mejores padres .

AGRADECIMIENTOS

A mis hermanos Antonio y Fernando, y a sus hermosas familias.

En recuerdo de mis abuelas María y Guillermina, formadores de nuestra familia.

A mi tía Joela por su cariño entrañable.

Con total aprecio al Dr. Ralph Algazi por su apoyo inestimable.

A nuestros amigos entrañables Cristy, Esperanzita, Joe y Bruce, que nos apoyaron en EUA.

A mi Facultad de Ingeniería, con aprecio a mis amigos académicos que me impulsaron a la realización de este trabajo.

A las personas tan estimadas de FI, DGIA, DGAPA y UCD que me apoyaron para mi estancia doctoral en la Universidad de California en Davis

Con especial afecto al Ing. José Manuel Covarrubias Solís por su amistad y ejemplo.

RESUMEN

El propósito de este trabajo es presentar una nueva metodología en la decimación de palabras basada en bandas críticas; una nueva técnica de parametrización de palabras basada en la transformada de Karhunen-Loève; una nueva técnica de detección de voz, un nuevo tipo de unidades básicas de procesamiento que se crean a partir de cambios espectrales fuertes en una palabra o frase y que tienen una correlación lingüística; un rediseño de las técnicas de clasificación de palabras: ajuste dinámico en el tiempo, cuantización vectorial, modelos ocultos de Markov y ajuste dinámico en el tiempo de un solo paso, las tres primeras para palabras aisladas y la última para palabras conectadas; y finalmente la aplicación de estas técnicas al reconocimiento de palabras aisladas y conectadas.

El aporte de este trabajo es el mejoramiento en la velocidad postprocesamiento con rendimientos comparables a los mejores sistemas de reconocimiento en similares condiciones, aspecto crítico para el éxito comercial de estos sistemas. También, se aporta un conjunto de experimentos sobre reconocimiento de voz en español hablado en México, hasta ahora tan precarios, pero tecnológicamente muy importantes.

Para palabras aisladas, se obtuvieron tasas de reconocimiento mayores a 99% para bases de datos de bajo ruido en inglés, y hasta de 98% en una base de datos con ruido alto en español. Los resultados de reconocimiento para palabras aisladas son comparables a los mejores obtenidos a nivel mundial. En el caso de palabras conectadas, los resultados son inferiores con respecto a los que se han obtenido para otros sistemas de reconocimiento conocidos con bases de dígitos conectados de bajo ruido en inglés, sin embargo, las bases de entrenamiento son mucho menores en nuestro caso.

ÍNDICE

TEMA PAG.

INTRODUCCIÓN

En la Feria Mundial de Nueva York de 1939, los laboratorios Bell sorprenden al mundo al mostrar el primer sintetizador de voz. En 1941, estos laboratorios diseñan el espectógrafo de voz usado en principio para fines militares, y desde 1944 para ayuda de personas sordas. En 1948 varios acontecimientos notables se producen en la ingeniería de comunicaciones, entre los que destaca la publicación del artículo de Claude Shanon "A mathematical theory of communication", el cual impulsa desde entonces una nueva área de las comunicaciones llamada procesamiento de señales. Esta área se plantea como uno de sus primeros objetivos es diseñar transductores automáticos voz-texto-voz.

Desde el primer sistema de reconocimiento de dígitos aislados de laboratorios Bell en 1952 hasta los hoy múltiples sistemas de reconocimiento de palabras continuas, sigue vigente el sueño de los años 40 de diseñar una máquina de escribir manejada por voz, en condiciones normales y con cualquier persona. Hasta hoy, los diferentes sistemas de reconocimiento continuo sólo son realmente confiables en condiciones de laboratorio, aunque sólo es cuestión de breve tiempo para obtener aplicaciones comerciales exitosas.

El propósito de este trabajo es presentar una nueva técnica de parametrización de palabras basada en la transformada de Karhunen-Loève (KLT), un nuevo tipo de unidades básicas de procesamiento, rediseñar algunas técnicas de clasificación, y finalmente aplicar lo anterior al reconocimiento de palabras aisladas y conectadas. El aporte de este trabajo es el mejoramiento en la velocidad postprocesamiento con rendimientos comparables a los mejores sistemas de reconocimiento, aspecto crítico para el éxito comercial de estos sistemas. También, se aporta un conjunto de experimentos sobre reconocimiento de voz en español hablado en México, hasta ahora tan precarios, pero tecnológicamente muy importantes.

Este trabajo se divide en un primer capítulo en donde se revisa el desarrollo histórico del área con una breve descripción de los sistemas de reconocimiento, así también se presentan las técnicas de análisis de mayor relevancia actual. A lo largo de los capítulos subsecuentes se describen los métodos utilizados en este trabajo. Así, en el segundo capítulo se presentan las técnicas de preprocesamiento, haciendo énfasis en el uso de bandas críticas que nos permiten una primera compresión de la señal. En el laboratorio de la Universidad de California en Davis (UCD), donde se desarrolla inicialmente el trabajo, se aplica inusualmente el concepto de bandas críticas al procesamiento de voz, de acuerdo con las propuestas del tutor

Al inicio del tercer capítulo se abordan algunos aspectos de la fonética acústica del español hablado en México, limitado a la región central, esto es muy importante porque son muy escasos en este país estudios al respecto. Posteriormente, se describe la técnica diseñada para las unidades de procesamiento, nuevamente se realiza aquí una contribución al utilizar unidades acústicas como unidades básicas del reconocimiento de voz. Para la obtención de subpalabras acústicas se diseñan dos métodos; uno basado en cocientes de máxima similitud, el cual se aplica también para detectar voz en presencia fuerte de ruido ambiental; y otro en las variaciones de los valores propios de la transformada de Karhunen-Loève.

En el cuarto capítulo se presentan las características esenciales de la transformada de Karhunen-Loève y cómo se aplica ésta al procesamiento de voz en tiempo corto. En este capítulo se presentan al inicio propiedades de esta transformada, posteriormente, se expone la estrategia de cálculo para voz, y finalmente se verifica su robustez para segmentos acústicos.

En el quinto capítulo de desarrollan las técnicas de clasificación para palabras aisladas y su desempeño. Se presentan las tres técnicas más utilizadas en la actualidad para este tipo de reconocimiento: alineamiento dinámico en el tiempo, cuantización vectorial y modelos ocultos de Markov, y se aplican tanto para bases de datos en el idioma inglés como en español.

En el sexto y último capítulo se describen las técnicas de clasificación aplicadas a palabras conectadas, esto es: alineamiento dinámico en el tiempo y modelos ocultos de Markov, se presentan también los métodos de entrenamiento más usuales y finalmente los resultados obtenidos al aplicar alineamiento dinámico en el tiempo a cadenas de dígitos en el idioma inglés.

En el anexo se mencionan algunos de los artículos publicados en memorias de congresos, resultado parcial de este trabajo.

Para palabras aisladas, se obtuvieron tasas de reconocimiento mayores a 99% para bases de datos de bajo ruido y hasta de 98% en bases de datos con ruido alto. Se utilizó la base datos de Texas Instruments que es un estándar a nivel mundial, tanto para palabras aisladas como para conectadas. Los resultados de reconocimiento para palabras aisladas son comparables a los mejores obtenidos a nivel mundial y para conectadas son menores con respecto a los que se han obtenido para otros sistemas de reconocimiento conocidos, pero con bases de datos de entrenamiento mucho menores.

El tiempo prolongado en que se desarrolló de este trabajo merece al menos una breve explicación. Así también se mencionan las contribuciones del autor a las técnicas de análisis y reconocimiento de voz presentadas, para deslindarlas de las investigaciones realizadas con anterioridad por el tutor con otros alumnos de doctorado, aspecto que se amplía en el capítulo primero.

Al iniciar la investigación doctoral en Ciudad Universitaria, bajo la dirección del Dr. Andrés Buzo en 1990, ésta se dirige a la aplicación de Modelos Ocultos de Markov (HMM) al reconocimiento de palabras aisladas . Se obtienen los primeros resultados para una base de datos muy pequeña en español de un solo hablante con precisiones bajas, de 88%. Esta investigación se ve suspendida meses después por la separación de la UNAM del tutor.

El autor decide continuar el trabajo de investigación doctoral, en la UCD bajo la tutela del Dr. Ralph Algazi, de 1991 a 1993. Éste ya había desarrollado con otros alumnos algunas aplicaciones originales de la KLT para codificación y reconocimiento de voz.

El autor contribuye con la introducción del filtro de preénfasis y modificaciones al segmentador acústico para variar el número de segmentos. También diseña una técnica basada en la variación de la energía contenida en los valores propios de la KLT, con mejores resultados. Para la fase de parametrización mediante la KLT, capítulo cuarto, se reformularon los procedimientos que utilizaron el tutor y el Dr. Irvine de UCD.

En la fase de clasificación, el tutor con estudiantes de UCD usaron el método DTW para reconocimiento de dígitos aislados en inglés con alta precisión. El primer trabajo del autor fue rediseñar la técnica DTW y mejorar los resultados de estos experimentos.

Posteriormente el autor desarrolla en México, desde 1994 y en tiempo parcial, los otros sistemas de reconocimiento de palabras aisladas y el de reconocimiento de palabras conectadas, para el idioma inglés. Se crea una base de palabras aisladas en español en condiciones de alto ruido ambiental, con la cual se realiza detección de voz y el reconocimiento de palabras aisladas con los tres métodos de clasificación

Capítulo 1

SISTEMAS DE RECONOCIMIENTO

El objetivo de los sistemas automáticos de reconocimiento de voz es contar con un transductor automático de voz continua a texto, solo es cuestión de tiempo para que se pueda contar con éste. La voz es el medio incuestionable de la comunicación humana, al ofrecer manos y ojos libres que otros medios de comunicación no tienen. No sólo se obtienen las ventajas inherentes a la comunicación humana sino también a la necesidad de comunicarse eficientemente con las máquinas, la comunicación por voz se caracteriza por su sencillez, conveniencia y rapidez.

La historia del procesamiento de voz data de fines de los años 40. Muchas técnicas han sido elaboradas desde entonces, hasta hoy se han comercializado sistemas de reconocimiento de palabras continuas, mucho de esto gracias al desarrollo del hardware en procesadores digitales de señales. Actualmente se estudian los sistemas de reconocimiento de palabras continuas en laboratorios para mejorar su robustez, sin embargo, parece que tardará aún algún tiempo antes de que las máquinas puedan reconocer palabras de una conversación común y corriente con muy altos niveles de precisión. Las complicaciones existen no sólo en el hardware, sino también para utilizar la información lingüística y contextual.

Los sistemas de reconocimiento se distinguen por el tipo de parametrización de las palabras, por la segmentación de éstas, y por el tipo de clasificación que realizan.

En una primera fase llamada preprocesamiento, se adecua la señal para la fase de análisis; consiste, en general, en un filtrado antitraslape y otro de preénfasis, para este sistema se añaden un cambiador de muestreo y un decimador en bandas críticas. La fase de análisis consiste en la caracterización de segmentos de voz por parámetros robustos y mínimos que la caracterizan y distinguen. Otra parte es cómo se escogen dichos segmentos de palabras, posteriormente cómo se agrupan y almacenan las palabras que sirven de referencia o comparación, etapa de entrenamiento, y finalmente cómo se realiza el reconocimiento de las palabras por identificar, esto último llamado etapa de clasificación.

En la figura 1.1 se muestra un diagrama típico de un sistema automático de reconocimiento de voz, en la que se observan las etapas anteriores. Es pertinente mencionar que en sistemas de reconocimiento de palabras continuas se añade una etapa indispensable en la clasificación para procesar la información lingüística.

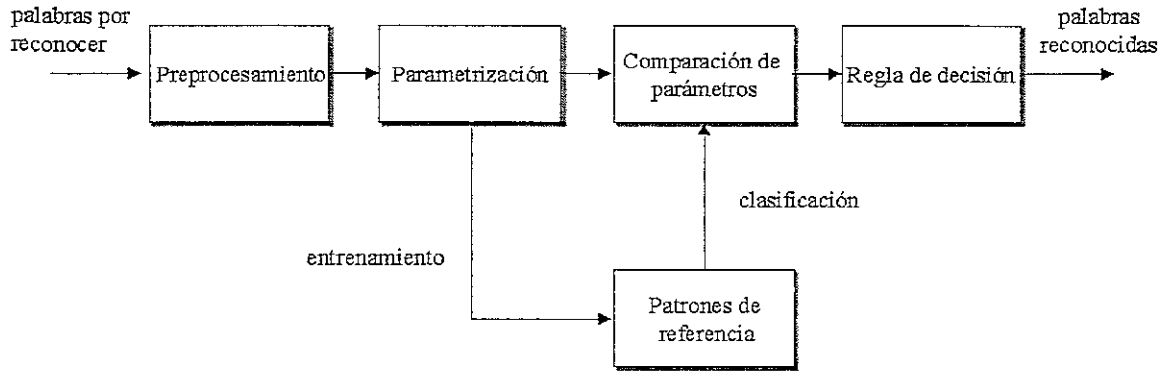


Figura 1.1. Sistema típico de reconocimiento automático de voz.

Existe un gran número de características que definen el tipo de reconocimiento, algunas de las más importantes son:

1. El tipo de palabras por reconocer. Los sistemas se aplican a palabras aisladas, conectadas y continuas. No existe una estandarización en los conceptos de estas dos últimas. Sin embargo, se considera que palabras continuas provienen de una conversación normal ya sea con un vocabulario limitado o ilimitado. Por palabras *conectadas se tiene una frase de longitud de palabras limitada proveniente de un* vocabulario también limitado, en este caso, no hay pausas en las cadenas de palabras. Existe otra aplicación consistente en la localización de palabras clave en frases.

2. Dependencia del hablante. Cuando se utilizan los mismos hablantes en el entrenamiento y en la clasificación, se denomina dependiente del hablante; si las personas son diferentes en estas dos fases, se denomina independiente del hablante. Obviamente, los sistemas dependientes del hablante tienen mejor desempeño.

3. Sistemas multiusarios o uniusuarios. Los sistemas uniusarios se utilizaron principalmente en los primeros sistemas de reconocimiento y actualmente para aplicaciones especiales. El número de hablantes es variable en las bases de datos, pero mayor a diez.

4. Sistemas adaptables o no adaptables. Los sistemas pueden ser adaptables a nuevos hablantes, modificando las bases de entrenamiento a nuevos usuarios o bien trabajar independientemente de los usuarios.

5. Tipo de unidad. Está determinada por el tipo de segmentación, existen tres orientaciones: matemática, lingüística y acústica. Se llaman así porque una palabra es segmentada en intervalos de tiempo aproximadamente iguales, sin considerar en su partición ninguna información lingüística o espectral de la señal. En las segmentaciones lingüísticas los segmentos corresponden a fonemas u otras unidades lingüísticas. La segmentación acústica fue utilizada poco en los 70 y abandonada por su variabilidad, en este trabajo se propone una más robusta.

6. Tipo de análisis. La voz se parametriza por segmentos muy breves de muestras de voz, llamados tramas. Como se ampliará más adelante en este capítulo, se utilizan parámetros que caracterizan de manera más robusta la voz en una trama que las propias muestras de voz. Sin embargo, estos vectores de parámetros se deben diseñar por segmentos mucho más cortos que la duración de un fonema. Las unidades de análisis (es decir, de parametrización) más utilizadas son los coeficientes de predicción lineal, LPC, y coeficientes cepstral, o variaciones de éstas.

En este capítulo se revisan primeramente los avances realizados mundialmente en el reconocimiento de voz, y posteriormente las técnicas de análisis actualmente más utilizadas, que constituyen la etapa de parametrización de voz, éstas son usadas universalmente en cualquier aplicación de procesamiento de voz. La técnica de parametrización utilizada en este trabajo se desarrolla en el capítulo cuarto.

1.1 Reseña histórica

Las primeras investigaciones sobre reconocimiento automático de voz se realizaron en la década de los 50. Sin embargo, podemos considerar que este campo de investigación nació una década antes con la invención del espectrógrafo de sonidos.

Los patrones visuales de voz fueron analizados desde 1941 por Potter, Kopp y Green, quienes lo utilizaron para fines militares y, posteriormente, en 1944 para ayuda a personas sordas[1]. Encontraron que los fonemas seguían patrones visuales de manera aproximada. Este análisis del espectro de voz fue la clave que abrió el interés sobre el reconocimiento de voz.

El problema de reconocimiento de palabras fluidas en un medio natural fue la meta principal en el área de reconocimiento de voz. Sin embargo, rápidamente se comprobó que esta labor era difícil de lograr, así que se dividió en etapas.

Lo primero fue reconocer palabras aisladas en diferentes grados de dificultad, desde el hablante dependiente al hablante independiente, iniciando con vocabularios pequeños hasta vocabularios extensos.

En 1952, Davis, Biddulph y Balashek desarrollaron en los laboratorios de la Compañía Telefónica Bell el primer sistema de reconocimiento[2]. Este sistema dividía el espectro de frecuencias en dos bandas que cortaban en 900 Hz y contaban el número de cruces por cero de la amplitud de ambas señales. Graficaban bidimensionalmente esta aproximación de los formantes F1-F2 y estos patrones eran correlacionados cruzadamente con diez patrones almacenados previamente, los cuales correspondían a los dígitos, con la mejor correlación se obtenía el resultado. Utilizando un solo hablante, el resultado fue un sorprendente 97% de precisión.

En 1958, Dudley y Balashek desarrollaron un sistema que usaba 10 bandas de frecuencia y derivaron información espectral que era comparada con patrones almacenados, incluyendo información sobre duración[3]. Una característica importante era que la segmentación de palabras se realizaba por unidades lingüísticas. Aunque la precisión en el reconocimiento era alta para un solo hablante, ésta disminuía drásticamente para más hablantes.

El uso de la computadora marcó el inicio propiamente de los sistemas de reconocimiento de voz. El primer sistema que utilizó una computadora digital fue en 1959[4]; en éste, Denes y Mathews incluyeron el concepto importante de normalización en el tiempo. Las palabras seleccionadas para pruebas de reconocimiento fueron comprimidas o extendidas por interpolación o decimación, de acuerdo con las señales entrenadas. Los resultados mostraron que los errores disminuían drásticamente al usar esta normalización, sobretodo para multihablantes.

En la década de los 60 se realizaron pruebas con hardware especial para sistemas de reconocimiento de palabras aisladas con vocabularios pequeños, entre los que destacaron el sistema de reconocimiento de IBM llamado *shoe recognizer* presentado en la Feria Mundial de 1962 y los sistemas para el idioma japonés, por Nagata en 1964, y alemán, por Musman y Steiner en 1965. El primer sistema de reconocimiento comercial apareció en 1972 desarrollado por las compañías Scope Electronics y Treshold Technology .

Los sistemas empezaron a ampliar el vocabulario en esa década, en 1966 Gold obtuvo 86% de precisión para 54 palabras y en 1968 Bobrow y Klatt lo mejoraron para este mismo vocabulario a 97%, considerando dos hablantes[5]. Su sistema se caracterizaba por utilizar un conjunto de elementos de una palabra (como segmentos vocales, estridencia, presencia de fonemas nasales, etc.) tomados en encendido-apagado y con su suma se clasificaba la señal.

En 1974, Martin y Grunza reportaron un sistema con una precisión de 99.79% para los 10 dígitos usando 10 hablantes y con 2 400 repeticiones de prueba[6]. Para un vocabulario de 34 palabras obtuvieron una precisión de 98.5% para 12 hablantes, 9 149 repeticiones en la prueba.

También en 1975, Itakura introdujo una nueva técnica de normalización llamada ajuste dinámico en el tiempo (DTW por sus siglas en inglés de *dynamic time warping*), su sistema obtuvo 99% de precisión para un vocabulario de 200 palabras geográficas en japonés con una SNR de 68 dB[7]. Para el alfabeto y los dígitos en inglés, que son más difíciles de identificar, obtuvo una precisión de 88.6%. White y Neely mejoraron las ideas de Itakura para obtener una tasa de precisión de 98% para este segundo vocabulario[8].

Los sistemas en general reducen su tasa de reconocimiento cuando son independientes del hablante. Scott. en 1975 y 1977, obtuvo para este tipo de sistemas tasas de precisión de

98% para los diez dígitos y cuatro comandos de control, 30 hablantes y 9 300 repeticiones en las pruebas; y una precisión de 96% para el mismo vocabulario pero 193 hablantes y 56 000 repeticiones[8][9].

Es natural que al incrementar el vocabulario por reconocer en un sistema de reconocimiento, su tasa de precisión disminuya. Por ejemplo, en el sistema propuesto por Cooler en 1977, se reportó una precisión de 99.9% para los diez dígitos con 20 000 repeticiones en las pruebas y 20 hablantes, pero con 5% de rechazo. Para el mismo sistema, usando 100 palabras, 10 hablantes y 100 000 repeticiones, la precisión se redujo a 95.7% con un rechazo de 5%[9].

Si bien en los años setenta se diseñaron sistemas para reconocimiento de palabras aisladas basados en DTW con vocabularios pequeños con una precisión alrededor de 99%, a finales de ésta década e inicios de los 80, se utilizaron métodos basados en agrupamientos de vectores por tramas cortas para este tipo de reconocimiento. Por cada cierto número de tramas, los vectores de parámetros LPC o cepstral se agruparon en conjuntos de centroides y se usó una suma de distancias, la más famosa la del tipo Itakura-Saito, para la etapa de clasificación, los resultados fueron equivalentes a DTW, pero se disminuyó el tiempo de reconocimiento de palabras[10][11]. Sin embargo, estos sistemas basados en agrupamientos fueron abandonados ya que al crecer el vocabulario perdían mucha precisión.

Un esfuerzo por aproximarse al reconocimiento de palabras continuas usando los resultados de palabras aisladas, lo realizó la IBM en 1985 al introducir el hardware de un sistema de reconocimiento de dictado, el cual reconocía oraciones pronunciadas con silencios. Este sistema usaba HMM , se entrenaron 100 oraciones con 1 107 palabras, se reconocieron oraciones con precisiones de 98% para oraciones pregrabadas, 96.9% para palabras leídas de un texto y 94.3% para palabras espontáneas[12].

Si el problema de reconocimiento de palabras aisladas ya había sido resuelto en gran medida en la década de los 80, la meta era obtener aproximaciones similares con vocabularios extensos y sistemas más rápidos. Se utilizó entonces la técnica de Modelos Ocultos de Markov (HMM), que si bien no logró los objetivos propuestos, su éxito fue su pertinencia para reconocimiento de palabras continuas. Cabe mencionar que algunos autores como Rabiner en 1989[13] reportan una pequeña baja en el reconocimiento de palabras aisladas usando HMM

comparado con DTW o VQ, otros autores lo confirman[9] y nuestros experimentos en la UNAM[14] aún más al usar modelos discretos.

El uso de redes neuronales (NN) en reconocimiento automático de voz se remonta a fines de la década de los 80. Esta técnica no produce mejores resultados que las anteriores y sí requiere mayor tiempo de procesamiento. Ha resultado adecuada en la adaptación de nuevos hablantes, ya que no requiere volver a entrenar todo el sistema como sucede con las técnicas mencionadas.

El segundo paso fue el reconocimiento de palabras conectadas. Difiere del reconocimiento de voz continua, porque en el reconocimiento de palabras conectadas la entrada es una secuencia de palabras de un vocabulario específico, y el reconocimiento se realiza con base en las combinaciones de unidades de referencia aisladas, generalmente, palabras. Por otra parte, en el reconocimiento de voz continua, la meta es reconocer un vocabulario ilimitado de cualquier hablante, y esto involucra el reconocimiento de unidades fonéticas. Sin embargo, ésta no es una distinción estándar; algunos sistemas se autodenominan reconocedores de voz continua porque utilizan unidades fonéticas a pesar de que usan un vocabulario limitado, como el sistema CMU-HARPY.

Los reconocedores de palabras aisladas y conectadas han tenido muchas aplicaciones. Principalmente se han aplicado a operadores telefónicos automáticos para llamadas asistidas, para información de aerolíneas y reservación de llamadas. Es importante resaltar que un problema intermedio que se ha popularizado en los últimos años es el reconocimiento de palabras clave dentro de frases habladas, que se llama palabras insertas, del inglés *word spotting*. Este problema ha emergido en aplicaciones comerciales para reconocedores de palabras aisladas cuando en muchos casos el hablante no contesta con una palabra específica sino con una frase; por ejemplo al llamar a la máquina operadora en vez de contestar "por cobrar", contesta, "por favor, quiero hacer una llamada por cobrar". Esta técnica puede ser considerada como un procedimiento de extracción de palabras en el reconocimiento de palabras conectadas. Existen muchos artículos acerca de las palabras insertas desde sus inicios en 1973[15] [16][17].

Los reconocedores de palabras conectadas comenzaron en 1970, para cadenas conocidas de tres dígitos, las precisiones se obtuvieron de 86.8% a 92%[6]. Un año más tarde, Sakoe y

Chiba aplicaron la técnica DTW a este tipo de reconocimiento[16]; sin embargo, el trabajo principal se realizó durante la década de los 80.

Las principales tareas para un exitoso reconocimiento de palabras conectadas son: identificar el número de unidades en una frase y dónde se localizan sus límites, el cálculo eficiente de las cadenas de palabras, y el realizar un procedimiento de entrenamiento adecuado.

Destacan los siguientes tres algoritmos que modifican los métodos DTW de palabras aisladas para cadenas de palabras, el primero por Sakoe, llamado de programación dinámica de dos niveles de empalme en 1979[19], posteriormente Myers y Rabiner en los laboratorios Bell proponen el algoritmo de edificio de niveles en 1981[20], que es menos adecuado que el anterior al utilizar más operaciones y resultados semejantes, el cual mejoran en 1989[21]; sin embargo, el método más rápido e igualmente preciso es el método de un solo paso propuesto por Ney en 1984[22], que a su vez se basó en resultados de Vintsyuk[23], y de Bridle y Brown[24].

Para mejorar la clasificación de palabras conectadas se han utilizado muchas técnicas de entrenamiento[9], entre las que destacan: el entrenamiento robusto[25][26], el entrenamiento entrelazado[27][28], de palabras clave[15][16][17], el entrenamiento por k-medias segmentado[29][30] y las técnicas de agrupamiento de HMM[31].

En la aplicación de HMM al reconocimiento de palabras conectadas, destacan los resultados de Rabiner en 1986, para cadenas de 2 a 7 dígitos usando entrenamiento k-medias segmentado , independiente del hablante con 50 hablantes, 525 cadenas para entrenamiento y el mismo número para clasificación, para obtener una precisión de 96% para longitudes conocidas y 94% para longitudes desconocidas[30]. El mismo autor, aplica posteriormente HMM por palabra en el entrenamiento, en un sistema independiente del hablante con 112 hablantes y un total de 8 565 cadenas, para obtener una precisión de 98.39% para longitudes conocidas y 97.07% para longitudes desconocidas[31]. Es ésta la más alta, o de las más altas precisiones reportadas para reconocimiento de palabras conectadas. Las precisiones que usan los métodos DTW con distintas técnicas de entrenamiento si bien son superiores a 95%[9][25][27] no sobrepasan las precisiones mencionadas.

El desarrollo de los tipos de reconocimiento de voz, aislado, conectado y continuo, como se ha observado no es linealmente sucesivo. En el caso del reconocimiento continuo, ya en 1966

Reddy realiza un reconocimiento de este tipo basado en fonemas, con precisiones muy bajas[32].

A fines de la década de los 80 e inicios de los 90, se publican los primeros resultados de reconocimiento de palabras continuas basados en HMM, destacan los sistemas HARPY y SPHINIX de la Universidad de Carnegie Mellon[33], el sistema BYBLOS de la BBN[34], el sistema del Lincoln Laboratory[35], del Stanford Research Institute[36], del Massachusetts Institute of Technology[37], y de los laboratorios Bell de ATT[38]. Cabe mencionar que estos sistemas fueron patrocinados por el Departamento de Defensa de los EUA, que seleccionó el sistema HARPY, el cual reconocía palabras continuas de un vocabulario de 1 011 palabras con una precisión de 95%, usando HMM con *biagramas* y el algoritmo de búsqueda de Viterbi[39].

En general, estos sistemas subsisten hasta hoy mejorados en el tipo de HMM, y en la búsqueda de unidades distintas a fonemas, entre las que destacan las *disílabas* (de dos sílabas) y los *triagramas* (formada por un fonema principal y parte de los fonemas antecesor y subsecuente, en tres modelos de Markov que conforman otro).

Estos sistemas utilizan reglas del lenguaje para disminuir las secuencias de búsqueda en la clasificación. Su importancia es tal que el propio campo ha modificado su nombre al de la Comprensión del Lenguaje Hablado (del inglés *Spoken Language Understanding*) , su diseño es crucial para la precisión y rapidez del sistema.

A la fecha, existen ya varios sistemas comerciales de reconocimiento de palabras continuas para cualquier hablante, con precisiones en condiciones naturales del orden de 90%[37], y sistemas para aplicaciones específicas como reservaciones en aerolíneas, de la empresa ATT, y conmutadores inteligentes, de la empresa "Telefonica". Así también, se encuentran ya desarrollados a nivel laboratorio sistemas de traducción simultánea[40][41]. En todos estos sistemas comerciales existe el objetivo de hacerlos más robustos y rápidos. Es aquí donde este trabajo adquiere su mayor importancia al proponer una nueva codificación de voz, y validar su robustez, rapidez y precisión con distintos tipos de métodos de clasificación de palabras aisladas y conectadas.

1.2 Estado actual en el análisis de voz

La señal de voz se procesa por tramas cortas cuya duración se establece de acuerdo con la propia variabilidad de la voz y con su característica cuasiestacionaria, entonces se utilizan ventanas de 128 o 256 muestras para inscribir en éstas las características del fonema a que correspondan. El tamaño óptimo de las tramas no es fácil de precisar, ya que por un lado deben de contener un período mínimo en fonemas sonoros, y por otro deben evitar contener cambios sustanciales en su espectro.

Las muestras de la señal de voz digitalizada en cada trama no constituyen parámetros robustos para la identificación de la trama o del fonema, se requiere de un conjunto de parámetros que la describan frecuencialmente por trama, más que su información temporal. El proceso de obtención de estos parámetros es llamado análisis de voz.

Los parámetros obtenidos serán parte fundamental para el subsecuente reconocimiento. Los parámetros usados casi de forma estándar son los coeficientes de predicción lineal, y los coeficientes cepstral. Ambos contienen información frecuencial de la trama, además de que comprimen la señal.

Los primeros representan los coeficientes de un filtro todo-polos que modela el tracto vocal durante esa trama, el filtro es variante en el tiempo. Los coeficientes señalan los formantes durante esa trama. En cambio los coeficientes cepstral son las amplitudes de las energías para diferentes frecuencias del logaritmo de la antitransformada de Fourier.

Aunque existe una gran variedad de parámetros que se derivan de los dos anteriores, los mencionados son los más populares por su robustez y facilidad en su cálculo, además de que caracterizan de manera precisa el comportamiento espectral en cada trama.

1.2.1 Análisis por predicción lineal

La idea básica en que se basan los códigos de predicción lineal (LPC) es que una muestra de voz $s(n)$, en el tiempo n, puede aproximarse por una combinación lineal de p muestras anteriores, esto es

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p s(n-p) = \sum_{i=1}^{p} a_i s(n-i) \qquad \text{Ec 1. 1}$$

donde los coeficientes a_i se consideran constantes en una trama, esto es, en un modelo de autoregresivo de predicción. Esta aproximación puede convertirse en una igualdad si incluimos el término de excitación $u_G(n)$, donde $u_G(n)$ es la excitación normalizada multiplicada por su ganancia, entonces

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + u_G(n), \qquad \text{Ec 1. 2}$$

esta ecuación se puede expresar como la función de transferencia

$$H(z) = \frac{S(z)}{U_G(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \qquad \text{Ec 1. 3}$$

Con relación al modelo digital de voz, $H(z)$ es el filtro todo polos que multiplicado por la excitación $u_G(n)$ produce la señal de voz $s(n)$. Los parámetros del tracto vocal del modelo mencionado están contenidos en los coeficientes del filtro variable.

El problema básico en la predicción lineal es determinar el conjunto de los coeficientes de predicción para cada trama que minimiza el error en la predicción, el cual está dado por

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i) \qquad \text{Ec 1. 4}$$

es importante anotar que los coeficientes mencionados deben estimarse por tramas cortas alrededor del punto n, dado el comportamiento no estacionario de la voz.

El objetivo es minimizar el error medio cuadrático para cierta trama n, que se denotará E_n, éste se define por

$$E_n = \sum_m e_n^2(m) = \sum_n [s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k)]^2 \qquad \text{Ec 1. 5}$$

Para obtener los predictores se deriva E_n respecto a cada coeficiente a_k, de donde,

$$\sum_{k=1}^{p} a_k \sum_m s_n(m-i) = \sum_m s_n(m) s_n(m-i) \qquad \text{Ec 1. 6}$$

donde $i = 1,2,\ldots,p$. Se forman ahora p ecuaciones con p incógnitas a_k. De esta ultima ecuación se puede identificar la covariancia de tiempo corto ϕ_n,

$$\phi_n(i,k) = \sum_m s_n(m-i) s_n(m-k) \qquad \text{Ec 1. 7}$$

entonces el error medio cuadrático mínimo está dado por

$$E_n = \sum_m s_n^2(m) - \sum_{k=1}^{p} s_n(m) s_n(m-k) = \phi_n(0,0) - \sum_{k=1}^{p} a_k \phi_n(0,k) \qquad \text{Ec 1. 8}$$

de esta ecuación se observa que el error mínimo consiste en un término fijo en el origen y términos que dependen de los coeficientes de predicción.

La solución de las p ecuaciones para a_k se puede determinar por dos métodos, de acuerdo con el criterio utilizado en el cálculo del error. Al sustituir la señal de voz multiplicada por una ventana, $s_n(m) = s(m+n)w(m)$ para $0 \le m \le N-1$ en la ecuación 1.5, se obtiene que las covariancias de tiempo corto ϕ_n, ecuación1.7, se reducen a los coeficientes de correlación

$$\phi_n(i,k) = r_n(j) = \sum_{m=0}^{N-1-j} s_n(m) s_n(m+j) \qquad \text{Ec 1. 9}$$

Si en cambio, a la señal no se le aplica una ventana y la sumatoria de la ecuación 1.5 se limita a la trama, las covariancias ϕ_n de la ecuación 1.7 se mantienen.

Ambos métodos conducen a errores en la obtención de las correlaciones; el segundo, a menores; sin embargo, el primero es más simple en su cálculo. Para el primer método se tiene un sistema de ecuaciones lineales donde la matriz de coeficientes es de Toeplitz y se aplica en su solución un método numérico llamado de Durbin. En el segundo método, la matriz de coeficientes es solo simétrica y el método numérico es llamado de Cholesky .

El cálculo de estos coeficientes para cada trama de longitud requiere la solución de sistemas p de ecuaciones y p incógnitas, donde p-1 es el orden de las autocorrelaciones. Las muestras de voz tienden a ser independientes rápidamente.

Existen varios métodos numéricos para la solución de los coeficientes LPC, entre los que destacan los de Blankenship y Kendall[42]; sin embargo, al reducir a menos de 10 el orden de los sistemas, no se aumenta significativamente el número de operaciones al usar el método más común de Durbin[42]. Este método consiste en los siguientes pasos:

1.Inicio.

$$E^{(0)} = R(0) \qquad \text{Ec 1. 10}$$

2. Primera iteración de $1 \le i \le p$

$$k_i = \frac{R(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R(i-j)}{E^{(i-1)}} \qquad \text{Ec 1. 11}$$

$$\alpha_i^{(i)} = k_i \qquad \text{Ec 1. 12}$$

3. Segunda iteración inmersa en la primera, de $1 \le j \le i-1$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \qquad \text{Ec 1. 13}$$

Regresa a la primera iteración,

$$E^{(i)} = (1 - k_i^2) E^{(i-1)} \qquad \text{Ec 1. 14}$$

$$a_i = \alpha_i^{(p)} \qquad \text{Ec 1. 15}$$

donde a_i, $i = 1,2,...,p$, son los coeficientes LPC de orden p.

De acuerdo con su interpretación física, los coeficientes LPC constituyen los coeficientes del filtro todo polos variante en el tiempo que caracteriza al tracto vocal. Su capacidad de modelar el espectro se muestra en la figura 1.2 obtenida por Markel[43], donde se muestra también la magnitud de la transformada rápida de Fourier (FFT), para una trama de 20 ms, donde la señal fue muestreada a 20 kHz con ventanas de Hamming, la trama está contenida en una trama del fonema en inglés /ae/, el orden del predictor es 28.

Se requieren un polo por kHz y de 3 a 4 polos para representar el espectro de excitación y de radiación, así para la banda de 4 kHz, ocho coeficientes son suficientes para caracterizar una trama de voz.

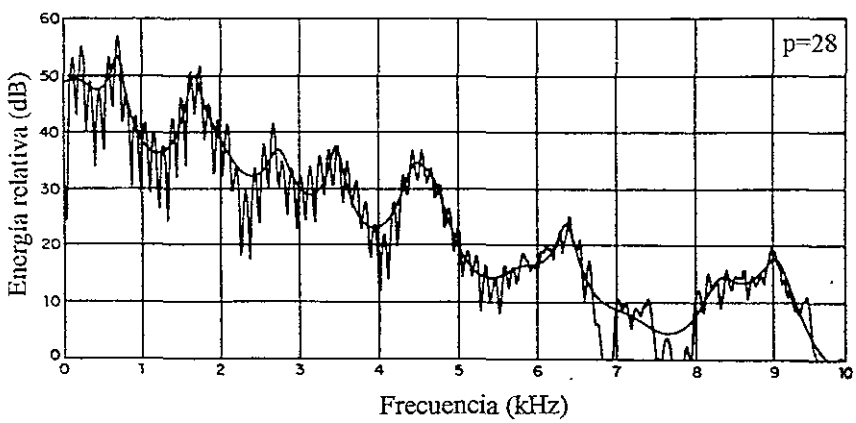


Figura 1.2. Magnitud de la FFT y coeficientes LPC de una trama del fonema en inglés /ae/[43].

La distancia que se deriva para comparar coeficientes LPC es la distancia de Itakura-Saito, de todas sus versiones la que se utiliza en este trabajo es la desarrollada por sus creadores[7] y consiste en

$$d(a_R a_T) = \log\left(\frac{a_R R_T a_R}{a_T R_T a_T}\right) \qquad \text{Ec 1. 16}$$

donde a_R es el vector LPC para una trama de referencia, a_T es el vector LPC para una trama de prueba y R_T es la matriz de autocorrelación de una trama de prueba. Se han desarrollado

otras variantes de esta distancia, por ejemplo para que sea una métrica, pero sin consecuencias significativas para el reconocimiento.

1.2.2 Análisis por coeficientes cepstral

Las señales de voz $s(n)$ pueden considerarse como la resultante de una modulación en amplitud, donde los pulsos glotales $p(n)$ que definen el tono representan la portadora mientras que el tracto vocal $h(n)$ representa la moduladora, esto es $s(n) = g(n) * h(n)$, la señal del tracto es una señal de menor frecuencia que la del pulso glotal.

Para las diferentes aplicaciones del procesamiento de voz, la señal del tracto vocal es mucho más importante que su portadora, ya que contiene los formantes y en general las características espectrales de cada fonema, de tal manera que al conocer esto desde la década de los 70, se buscó separar la señal del tracto $h(n)$ de la señal de voz $s(n)$. Desde el punto de vista teórico, esta separación se podría obtener al aplicar primeramente el logaritmo al espectro de la señal, posteriormente la transformada inversa de Fourier a $\log\{|S(w)|\}$, después un filtro paso bajas y finalmente la transformada de Fourier al resultado anterior, para que de acuerdo a la propiedad logarítmica,

$$\log\{|S(\omega)|\} = \log\{|P(\omega)|\} + \log\{|H(w)|\} \qquad \text{Ec 1. 17}$$

y a los comportamientos espectrales disjuntos de $P(w)$ y de $H(w)$ se obtiene el espectro logarítmico $\log\{|H(w)|\}$ del tracto vocal[43].

De manera más práctica este resultado se puede obtener al suavizar el espectro $\log\{|S(w)|\}$, o bien a partir de los coeficientes LPC[42], en este último caso son llamados coeficientes LPC cepstral, que se obtienen cuando el espectro de $\log\{|S(w)|\}$ es de fase mínima. De la ecuación 1.15, se proseguiría de la siguiente forma:

1. Inicio,

$$c_0 = \ln \sigma^2 \qquad \text{Ec 1. 18}$$

2. Primera iteración, para $1 \leq m \leq p$, donde p son los coeficientes LPC considerados,

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} \qquad \text{Ec 1. 19}$$

3. Iteración, para $m > p$, independiente de la anterior,

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} \qquad \text{Ec 1. 20}$$

Los vectores cepstral requieren solo de la distancia euclidiana para medir la distancia entre estos vectores por trama.

Los coeficientes cepstral son más robustos que los coeficientes LPC, son los más utilizados y en conjunto con los coeficientes LPC y PARCOR ocupan el universo de los parámetros utilizados en todas las aplicaciones de los sistemas de procesamiento de voz.

1.3 Sistemas de reconocimiento propuestos

A continuación se describen los diferentes sistemas de reconocimiento de palabras aisladas y conectadas propuestos. En particular se desarrolla un sistema de análisis de palabras basado en la transformada de Karkunen-Loève; tres sistemas de clasificación de palabras aisladas basados en Ajuste Dinámico en el Tiempo (DTW), Cuantización Vectorial (VQ) y Modelos Ocultos de Markov; y un sistema de clasificación de palabras conectadas basado en DTW de Un Solo Paso.

En la figura 1.3 se muestra un diagrama de bloque del sistema DTW. En los siguientes sistemas de clasificación variarán los tres últimos bloques y se mantendrán intactos los dos primeros.

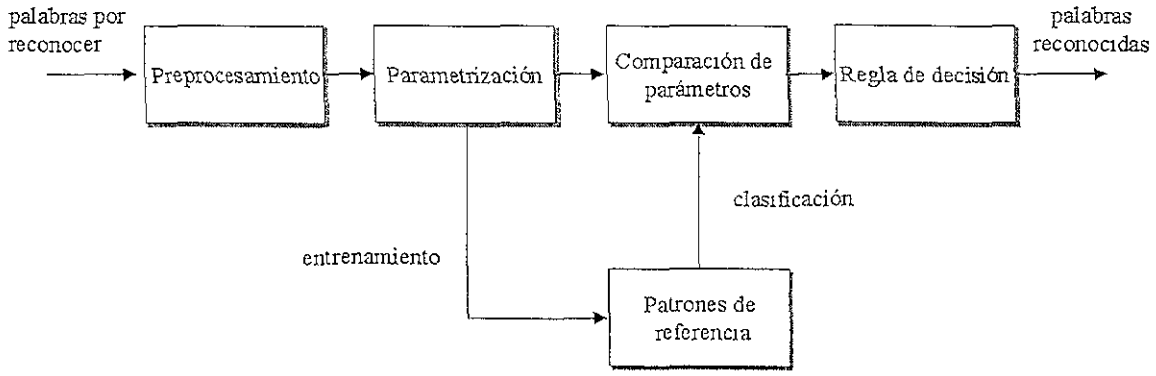


Figura 1.3. Diagrama de bloques del sistema de reconocimiento de palabras aisladas usando DTW.

En la figura 1.3 puede observarse una primera fase de preprocesamiento. Esta se desglosa en la figura 1.4 y se desarrolla en el capítulo siguiente. En los bloques, el autor añade un filtro de preénfasis para compensar la baja energía de los fonemas fricativos y el error consecuente en la predicción lineal.

El modificador de muestreo desarrollado en la UCD simplifica algunas observaciones; al obtener muestras de 10 000 Hz comparado con la base de datos en inglés, de Texas Instruments, que muestrea a 11 025 Hz y la tarjeta Sound Blaster, utilizada para la base de datos en español, que muestrea también a 11 025 Hz.

La parte más importante de la fase de preprocesamiento es el decimador en bandas críticas. De éste se obtienen vectores de dimensión 1x18 por trama, que constituyen una compresión espectral muy importante. Comparada con la codificación LPC o cepstral se aumenta la dimensión al doble, pero su procesamiento es mucho más rápido[44]. Aquí no hubo cambios sustanciales en las ideas ya desarrolladas en la UCD, pero sí observar que se puede reducir el uso a 16 bandas sin pérdida en la precisión del reconocimiento.

Es importante mencionar que esta decimación puede obtenerse de un proceso de filtrado. En esta investigación la decimación se obtiene a través de la transformada rápida de Fourier (STFT), ya que se pretendió que fuera un sistema compatible para codificación y reconocimiento. De hecho el autor toma el sistema de codificación, y adapta los bloques de preprocesamiento y parametrización KLT para reconocimiento, añadiendo la etapa de clasificación.

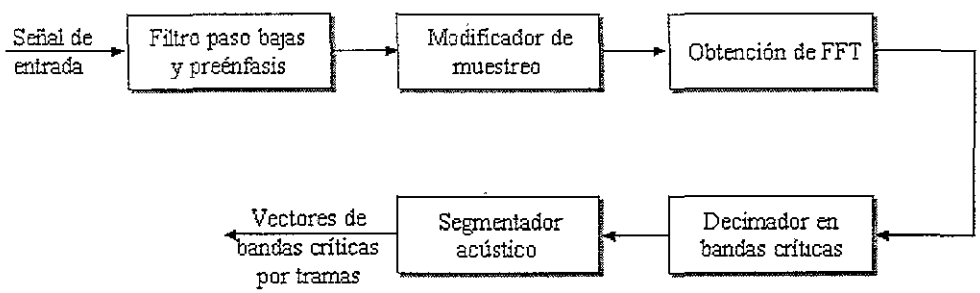


Figura 1.4. Diagrama de bloques de la fase de preprocesamiento de los sistemas de reconocimiento

El *segmentador acústico está basado en pruebas de cocientes de máxima similitud* (MLR), y se desarrolla en el capítulo 3. El segmentador fue modificado por el autor para permitir fijar libremente el número de segmentos acústicos. También se utilizaron otras técnicas, como la energía contenida en los valores propios de la KLT para fijar simultáneamente los segmentos acústicos. La fase de parametrización por la KLT se desarrolla en el cuarto capítulo, donde se utilizaron los procedimientos del tutor y del Dr. Irvine de UCD[45].

Las fases de clasificación ya son diseño del autor, las tres técnicas de clasificación de palabras aisladas se presentan en el capítulo 5. Para DTW se menciona el diseño original de trayectorias locales y globales que proporcionan más rapidez y mejoran su rendimiento[46]. Así también, las repeticiones de entrenamiento pueden ser las mismas que para clasificación, a

excepción obviamente de la de prueba. Los vectores espectrales por trama o segmento acústico de cada repetición se almacenan en el entrenamiento.

En la figura 1.5 se muestra el diagrama de bloques para el caso de clasificación por VQ. A diferencia de DTW, en este caso la segmentación acústica es fija para todas las repeticiones, y las repeticiones de entrenamiento deben ser distintas a las de clasificación.

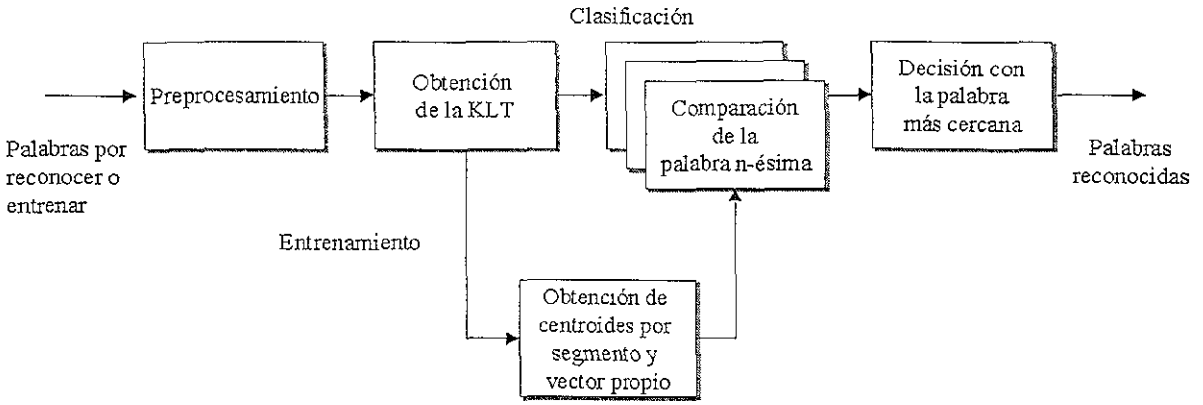


Figura 1.5. Diagrama de bloques del sistema de reconocimiento de palabras aisladas usando VQ.

El algoritmo de agrupamiento usado es K-medias modificado tanto para entrenamiento como para clasificación. Se utiliza un agrupamiento por cada segmento acústico, es decir, un VQ multiseccionado.

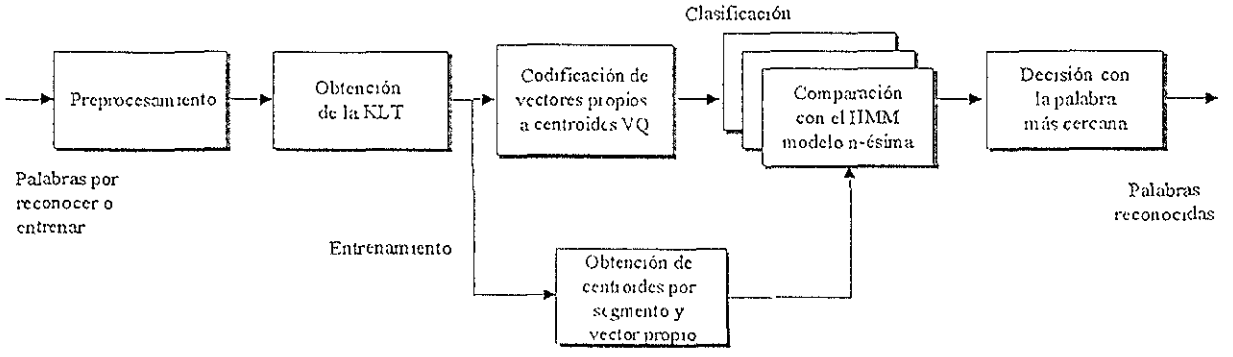


Figura 1 6 Diagrama de bloques del sistema de reconocimiento de palabras aisladas usando HMM

En la figura 1.6 se muestra el diagrama de bloques para el caso de clasificación por HMM. En este sistema, con las palabras de entrenamiento se diseña un modelo discreto de Markov por palabra, y con repeticiones distintas se realizan las pruebas

En este caso, se utiliza el algoritmo de Viterbi en la clasificación de palabras. Para los tres sistemas se utilizaron bases de datos en inglés y en español. Se compararon con los sistemas más conocidos que utilizan coeficientes LPC y cepstrals.

La clasificación de palabras conectadas se presenta en el capítulo seis y último. El algoritmo DTW es llamado de un solo paso, y se aplicó para cadenas de dígitos en inglés, ver figura 1.7.

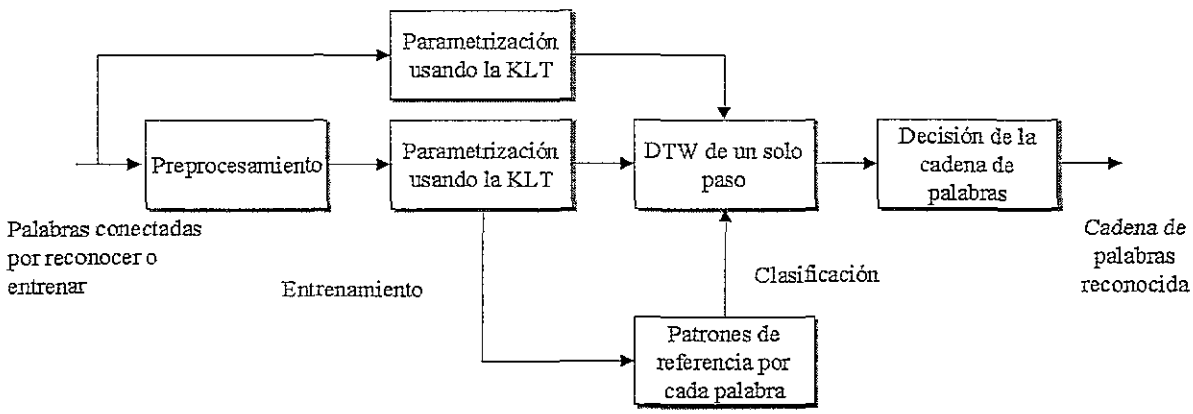


Figura 1.7. Diagrama de bloques del sistema de reconocimiento de palabras conectadas usando DTW de un solo paso.

En la figura 1.7 se observan los dos modos de clasificación para este sistema, con la longitud de la cadena de dígitos conocida y el caso contrario; este último es de menor precisión. La precisión en la clasificación de cadenas de dígitos se presenta por número de palabras identificadas en el orden en que deben aparecer, y de la cadena en su totalidad. Como se describirá en el capítulo seis, la precisión aumenta al utilizar en el entrenamiento palabras provenientes de cadenas de dígitos.

Capítulo 2

PREPROCESAMIENTO

> Nada en el sistema auditivo y en la psicoacústica es más común que las bandas críticas. Aparece en los estudios del tono, sonoridad, inteligibilidad, enmascaramiento, fatiga, percepción de fase. Y finalmente, de un modo u otro, es parte de nuestro entendimiento final de cómo y por qué percibimos lo que llega a nuestros oídos. Jerry V. Tobias[47].

En la figura 2.1 se muestra un diagrama de bloques del preprocesamiento realizado en este trabajo. Básicamente consta de filtros antitraslape y preénfasis, un cambiador de muestreo, la aplicación de la transformada rápida de Fourier, un decimador en bandas críticas, y un segmentador acústico. Estos bloques se tratarán en este capítulo, a excepción del segmentador acústico que se tratará en el siguiente capítulo.

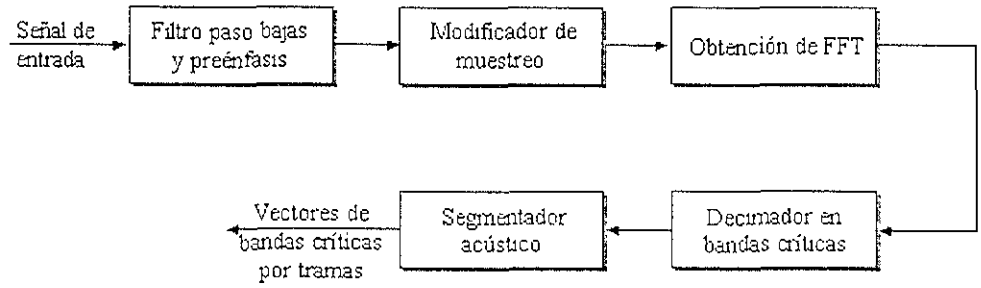


Figura 2.1 Diagrama de bloques de la fase de preprocesamiento del sistema de reconocimiento.

Cabe mencionar que los tres primeros bloques se tratarán brevemente, y el cuarto que se refiere al decimador de bandas críticas de manera extensa, en virtud de que es en éste donde se plantean innovaciones al procesamiento de voz tradicional.

2.1 Filtrado

Las señales grabadas tienen una frecuencia de corte específico, como son los archivos de palabras de *Texas Instruments* para palabras aisladas y conectadas, de 4 kHz. Sin embargo, para evitar errores que puedan conducir a traslapes por señales espurias arriba de esta frecuencia, es conveniente colocar un filtro pasobajas con frecuencia de corte de 4 kHz.

Se escogió un filtro tipo Chebychev de orden 8. Dado que no requerimos sintetizar la señal, se prefiere un filtro simple, de cortes suaves en ambos extremos del corte que otro de mayor pendiente, pero que pudiera introducir ruido en armónicas. En la figura 2.2 se muestra el espectro antes y después de aplicar este filtro, dado que la frecuencia original es de 11 025 Hz, en la gráfica de la derecha se observa en la región superior la energía casi nula, Como se usó una rutina de Matlab, la frecuencia está acotada superiormente a la mitad de la frecuencia máxima de la señal.

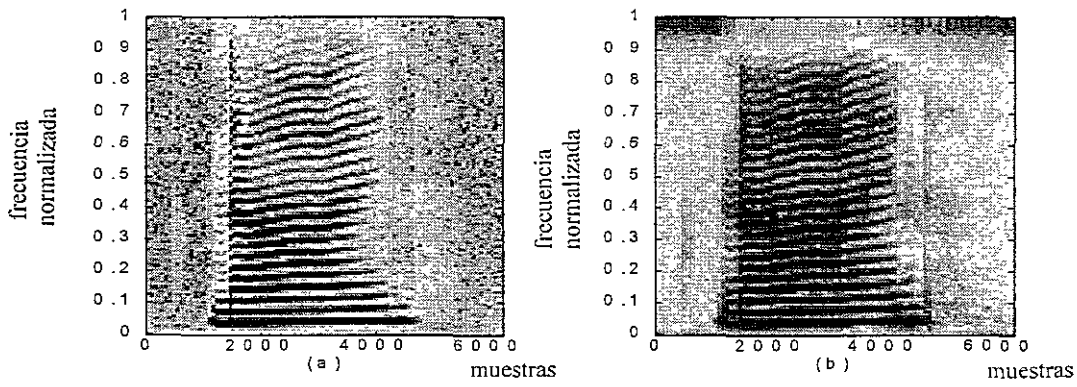


Figura 2.2. Espectrogramas de la palabra "nine", (a) antes y , (b) después de aplicar un filtro antitraslape .

En la generación del habla existe una tendencia pasobajas de -6 dB/octava. Esto es resultado de la combinación de -12 dB/octava debida a la fuente de excitación de voz y de 6 dB/octava debido a la radiación de la voz en la cavidad bucal. Es deseable entonces compensar con un aumento de 6 dB/octava en un rango apropiado. A esta compensación se le denomina filtro de preénfasis.

Adicionalmente, el filtro de preénfasis compensa las diferencias de energía entre fonemas sonoros y sordos fricativos, no sólo para observarlos adecuadamente, mediante los espectrogramas, sino para la obtención de valores significativos por trama o subpalabra de la

señal. En general, los fonemas sordos fricativos contienen su mayor energía en frecuencias altas, es decir entre 2 y 4 kHz; sin embargo ésta es aún muy baja comparada con la de fonemas vocálicos, por lo que un simple filtro pasoaltas como el de preénfasis permite obtener el resultado deseado. Su función de transferencia se puede establecer como:

$$H(z) = 1 - az^{-1} \qquad 0.9 \le a \le 1.0 \qquad \text{Ec 2. 1}$$

el valor más común del coeficiente a es 0.9.

En la figura 2.3 se muestra la aplicación de este filtro a una palabra aislada. En su espectrograma, figura 2.3(d), puede observarse el refuerzo de las frecuencias altas, que en el tiempo, figura 2.3(b), se traduce particularmente en la mayor amplitud de sonidos sordos.

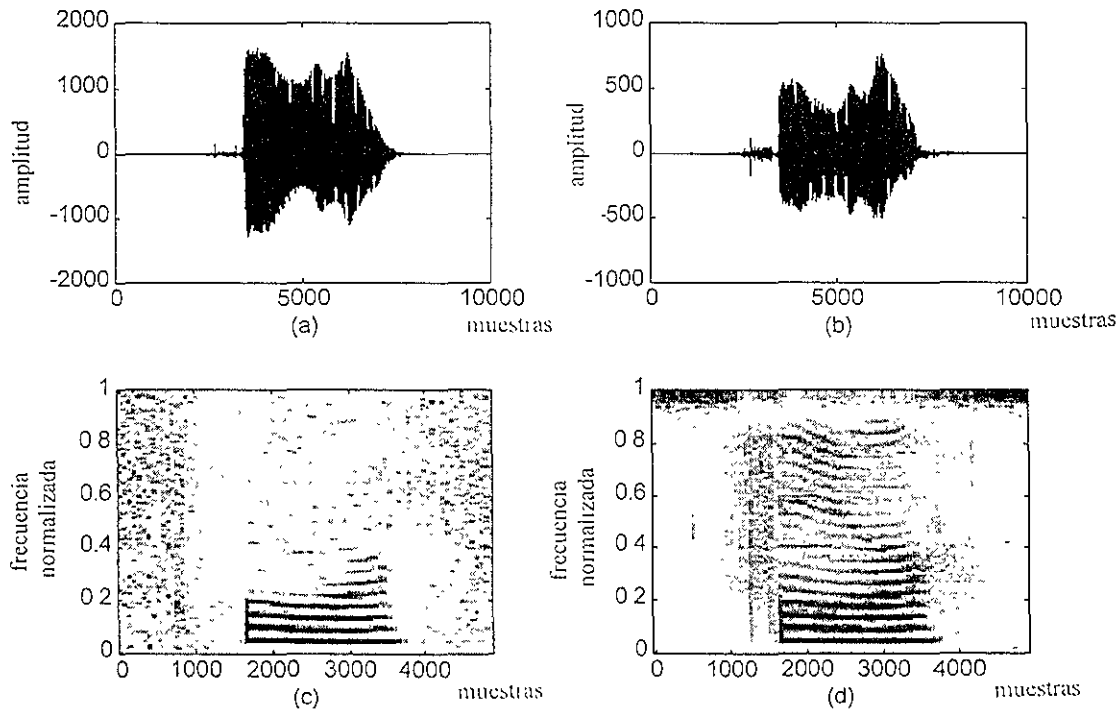


Figura 2.3. Gráficas en el tiempo (a) y espectrogramas (c) de la palabra "four", y estas gráficas (b) y (d) después de aplicar un filtro de preénfasis.

2.2 Modificador de frecuencia de muestreo

Al realizar la adquisición de señales de voz, se plantea el problema de las distintas tasas de muestreo a las que han sido discretizadas o grabadas, Es esencial que el procesamiento se realice con señales muestreadas a la misma tasa, ya que muchas de las comparaciones analíticas o gráficas que se realizan en el reconocimiento son temporales.

Al manejar siempre tasas de muestreo racionales, los cambios de tasas involucran un cambio de tasa de un valor racional V, que al expresarlo como el cociente de dos enteros, $V = ID$, nos denota que se debe realizar una interpolación de valor I, y posteriormente una decimación de valor D, o bien múltiplos de estos valores I y D. Los dos procesos deben realizarse en este orden para evitar problemas de traslape.

La interpolación más simple se da por inclusión de ceros a una señal discreta $s(n)$, resulta en la señal

$$s_I(n) = \begin{cases} s(n) & \text{para} \quad n = 0, \pm I, \pm 2I, \pm 3I, \pm 4I \\ 0 & \text{caso contrario} \end{cases} \qquad \text{Ec 2. 2}$$

cuyo espectro es una versión escalada en frecuencia de la señal original, esto es:

$$S_I(f) = \sum^{\infty} s_I(m) e^{-j2\pi fm} = \sum_{m=-\infty}^{\infty} s(m) e^{-j2\pi fmI} = S(If) \qquad \text{Ec 2. 3}$$

Si bien desde el punto espectral se obtiene un resultado óptimo, desde el temporal la distorsión introducida a la señal es inaceptable.

La interpolación lineal proporciona una señal temporalmente muy semejante a la original y la distorsión introducida al espectro es mínima, esta última depende de la amplitud de la señal y de su tasa de muestreo. Para señales de voz, que se muestrean arriba de 10 kHz, es suficiente con una interpolación de 5 puntos de pendiente unitaria.

2.3 Aplicación de la FFT

La FFT se aplica a ventanas de 256 muestras con un traslape de 20 muestras. Las ventanas son de tipo Hanning, la aplicación de ventanas se establece en el tiempo de acuerdo con la ecuación 2.4, donde M es el tamaño de la ventana.

$$w(n) = \begin{cases} 0.5 - 0.5\cos(2\pi n / M) & \text{para} \quad 0 \le n < M, \\ 0 & \text{caso contrario} \end{cases} \qquad \text{Ec 2. 4}$$

Este tamaño de ventanas es prácticamente un estándar en procesamiento de voz, junto con ventanas de 128 muestras. El objetivo es que contemple al menos un periodo de la señal, lo cual tampoco es trivial, ya que éste cuando existe es variable.

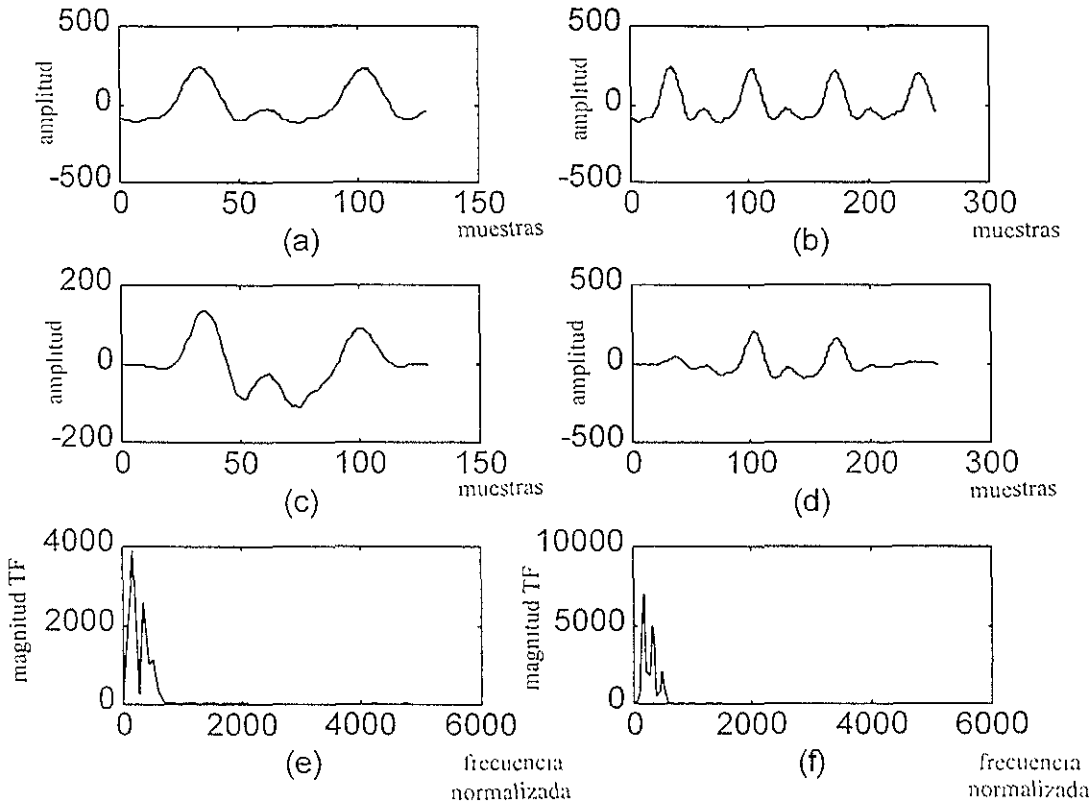


Figura 2.4. En el inciso (a) se presenta trama de 128 muestras en el tiempo, después la trama al aplicarle una ventana de Hanning (c), y finalmente la magnitud de la TF (e), en la derecha lo equivalente con una trama de 256 muestras (b), (d) y (f) La trama pertenece al fonema vocálico /u/.

Sin embargo, ambos tamaños de ventanas son lo suficientemente grandes para incluir prácticamente a todos los casos, como se observa en la figura 2.4(a) y 2.4(b). Así en la figura 2.4(a) se observa que para una trama del fonema vocálico /u/, el cual es cuasiperiódico, con una longitud de 128 muestras se capta un periodo y medio, y con una longitud de 256 muestras se incluyen tres periodos, en la figura 2.4(b). Por otro lado, se pretende que las ventanas sean lo suficientemente pequeñas para que las tramas sean estacionarias o levemente cuasiestacionarias.

El tipo de ventana se escogió para que la etapa de codificación del sistema, no contemplada en este trabajo, fuera más sencilla de realizar. En las figuras 2.4(c) y 2.4(d) se observa el efecto de este tipo de ventanas sobre una trama, donde los valores extremos de las tramas son reducidos significativamente. Al realizar un traslape, en este caso de 20 muestras, se compensa este efecto.

Se trabaja únicamente con la magnitud de la FFT, ya que como se demostró desde el siglo anterior por Helmholtz, los cambios de fase no son detectables, hasta cierto punto, por el oído y sus cambios son poco relevantes en el procesamiento de voz. En las figuras 2.4(e) y 2.4(f) se muestra la magnitud de la FFT para la trama mencionada anteriormente, como se espera resaltan los formantes del fonema y el mayor contenido de energía en bajas frecuencias.

2.4 Decimador en bandas críticas

Se llama enmascaramiento al fenómeno en el que un sonido interfiere con la percepción de otro. Esta interferencia depende tanto de la intensidad como de la frecuencia de ambos sonidos. Un tono enmascara más a otro cuando su frecuencia es cercana a este último.

Fletcher en 1937[47] observó que cuando un tono puro era enmascarado por ruido de banda ancha sólo una banda pequeña centrada alrededor del tono contribuía a tal efecto, la llamó banda crítica. Desde entonces, las bandas críticas han aparecido en un gran número de fenómenos perceptuales, en los que se incluyen: el enmascaramiento por un segundo tono, la sensibilidad a la modulación en fase, y la percepción de disonancias musicales, entre otros.

La utilización de las bandas críticas en el área ha sido muy reducida. Destaca su empleo en cancelación de ruido por Peterson y Boll en 1981 los cuales usaron substracción de espectros, y la aplicación de Jayant en 1995 en filtros para codificación. No han sido utilizadas en el reconocimiento automático de palabras.

2.4.1 Enmascaramiento

El enmascaramiento se puede clasificar en dos tipos. Cuando los sonidos son simultáneos pero de distinta frecuencia se llama enmascaramiento en frecuencia; cuando un sonido está retrasado respecto a otro de igual o distinta frecuencia se llama enmascaramiento temporal.

Por el tipo de señal a la que se aplica tanto la señal que enmascara como a la señal enmascarada o a enmascarar, el enmascaramiento puede ser de tonos, ruido o palabras, o bien, combinaciones de éstas. Históricamente en este orden ha sido estudiado.

Primeramente se describirá el enmascaramiento en frecuencia entre tonos. El enmascaramiento en frecuencia de un tono respecto a otro es una función de su separación en frecuencia. Ya en 1876, Mayer realizó algunos experimentos de este tipo de enmascaramiento[48], y apuntó el comportamiento asimétrico del enmascaramiento que se puede observar en la figura 2.5 obtenida por Fletcher[49], pero ya referida por Wegel y Lane de los laboratorios Bell en 1924[48].

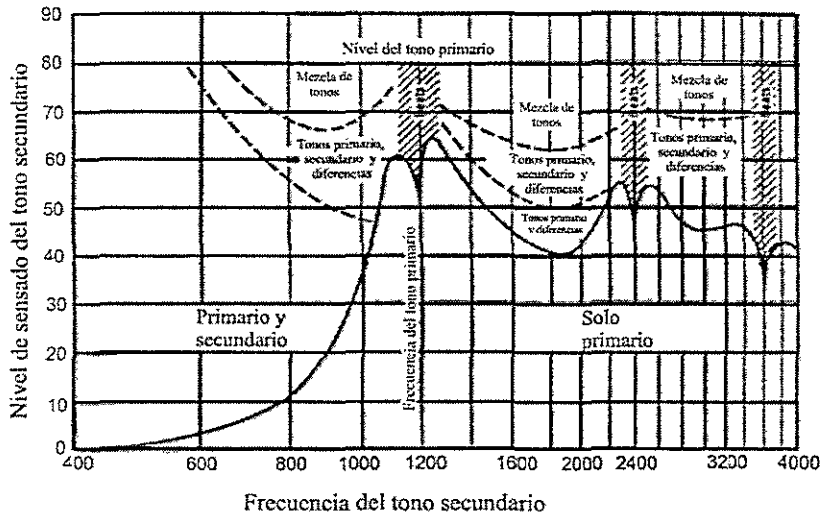


Figura 2.5. Enmascaramiento de tonos puros. El tono enmascarador de varias intensidades se fija en cada gráfica y las curvas indican el umbral de los tonos enmascarados[48].

En la figura 2.5 se puede observar, por ejemplo, que si un tono enmascarador se fija a 1 200 Hz y 80 dB, un segundo tono de 800 Hz puede oirse desde amplitudes de 12 dB. Sin embargo, cuando el segundo tono está dentro de 100 Hz de 1 200 Hz, el umbral de enmascaramiento es de 50 dB, este efecto ocurre también en altas frecuencias. Aún más, el umbral de enmascaramiento puede igualar al umbral en silencio cuando las frecuencias de ambos tonos están muy alejadas entre sí. También puede observarse que el enmascaramiento es mayor cuando el tono enmascarador tiene una frecuencia menor al tono enmascarado, que cuando su frecuencia es mayor al tono por enmascarar.

En las gráficas destacan los valles cuando el tono enmascarado se acerca a la frecuencia del tono enmascarador, o bien en las armónicas; estos se deben a distorsiones producidas por combinaciones de los dos tonos. Es importante mencionar que justo encima del umbral de enmascaramiento no solo se perciben los dos tonos sino también su diferencia cuando la frecuencia del tono enmascarado no es tan baja respecto al otro tono. Para intensidades mayores del tono enmascarado, se perciben los dos tonos y las combinaciones positivas de $mf_1 \pm n(f_1 - f_2)$ Hz, donde m y n son enteros dentro de cierto rango alrededor de la frecuencia del tono enmascarador f_1.

Acerca del enmascaramiento de tonos en presencia de una banda de ruido, vale la pena describir los experimentos de Hawkins y Stevens[48] en 1950, acerca del enmascaramiento de tonos por ruido blanco, llamado de banda ancha porque cubre todo el rango audible. Obtuvieron las amplitudes requeridas para detectar un tono por encima de la amplitud requerida por el mismo tono en ausencia de ruido, para tonos de 350, 500, 1 000, 2 800, 4 000 y 5 600 Hz, comparando para diferentes niveles efectivos (niveles totales) del ruido blanco. Este se consideró como $I + 10\log W$, donde I es el nivel espectral (niveles de densidad), esto es la potencia del ruido blanco medido en una banda de 1 Hz, y W es el ancho de banda en Hz del ruido que contribuye efectivamente al enmascaramiento (banda crítica). Así, por ejemplo, para un tono de 1 kHz y un nivel efectivo de 54 dB de ruido, el tono requiere aproximadamente 52 dB más para ser detectado que cuando se detectara en silencio.

Zwicker y Fasti obtuvieron resultados semejantes[50], pero ampliaron los experimentos a un gran número de tonos, figura 2.6. En esta gráfica, los umbrales de enmascaramiento de los tonos están considerados también para distintos niveles espectrales. Así para un tono de 2 kHz, el umbral es de aproximadamente 30 dB, cuando el nivel espectral es de 10 dB.

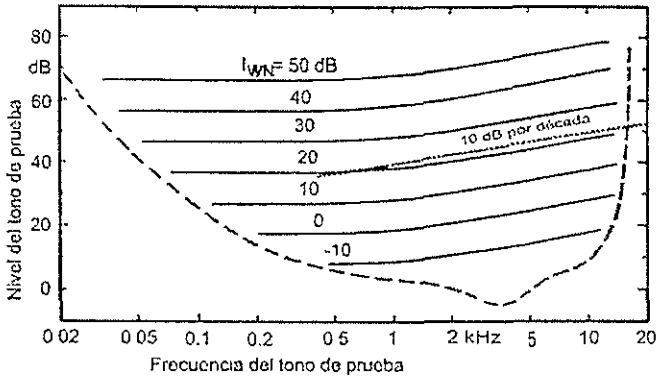


Figura 2.6. Enmascaramiento de tonos por ruido blanco de banda ancha de distintas intensidades, en función de la frecuencia de los tonos[50].

Otro experimento importante realizado por Zwicker y Fasti es para ruido blanco en una banda crítica o menor a ésta, en la gráfica 2.7 se puede observar que el umbral de enmascaramiento para tonos de 250,1 000 y 4 000 Hz es un poco menor al nivel espectral de 60 dB. Por otra parte, el umbral de enmascaramiento es casi igual a su nivel en presencia de ruido blanco de banda ancha, lo que verifica la tesis de que sólo el ruido dentro de una banda

crítica contribuye efectivamente al enmascaramiento, y también cómo el umbral decae muy fuertemente al alejarse de las frecuencias centrales de las bandas.

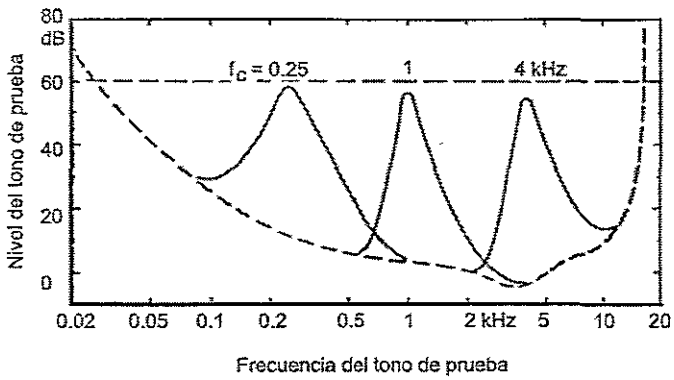


Figura 2.7. Enmascaramiento de tonos por ruido blanco en bandas críticas de distintas intensidades[50][50].

El enmascaramiento en tiempo ocurre cuando el sonido por enmascarar se presenta antes o después del sonido enmascarador, o bien cuando se presenta simultáneamente pero tiene una duración menor. Este fenómeno ha sido menos estudiado que el enmascaramiento en frecuencia, sin embargo es importante en muchas aplicaciones como la cancelación de reverberación.

Se define enmascaramiento hacia atrás o preenmascaramiento cuando el sonido enmascarador ocurre después del sonido a enmascarar, lo inverso se denomina enmascaramiento hacia adelante o postenmascaramiento. Estos enmascaramientos dependen entonces de la intensidad, frecuencia y tiempo de ocurrencia de ambos sonidos. Siempre el enmascaramiento hacia adelante tiene mucho mayor duración que el preenmascaramiento, el primero tiene lapsos de hasta 200 ms, y el segundo no más de 20 ms por lo que es frecuentemente ignorado.

Para el enmascaramiento hacia adelante, Lüscher y Zwislocki mostraron en 1949 que el enmascaramiento es función de la intensidad del sonido enmascarador, y disminuye rápidamente cuando el intervalo de tiempo entre sonidos aumenta[48][48]. Por ejemplo, para un tono de enmascarador de 80 dB y 400 ms de duración, el enmascaramiento es de 40 dB cuando la señal ocurre 20 ms después; el enmascaramiento cae a 0 dB cuando el intervalo se incrementa a 200 ms, esta caída es lineal con el tiempo.

La linealidad obtenida para el enmascaramiento hacia adelante no se produce para el precedente. En 1962, Elliot mostró que, por ejemplo, cuando el ruido empezaba 1 ms después de la señal, el enmascaramiento era de 60 dB, y decrecía a 20dB cuando el intervalo entre ambas se incrementaba a 10 ms, y finalmente era nulo para intervalos mayores a 25 ms[48].

En el caso de enmascaramiento simultáneo, éste depende de la duración del tono enmascarado en un ruido de duración larga. Así, entre mayor es la duración del tono menor es la del umbral de enmascaramiento. El comportamiento es lineal con una pendiente de 10 dB por octava, para ruido blanco[50].

2.4.2 Bandas críticas

El término bandas críticas se ha utilizado en un gran número de fenómenos empíricos donde la percepción de sonidos depende del ancho de banda de estos.

Como ya se mencionó, Fletcher acuñó el término bandas críticas en experimentos sobre enmascaramiento, señalados en el inciso anterior, y semejantes a los realizados por Hawking y Stevens, se considera que son insuficientes para determinarlas de un modo preciso. Muchos autores posteriores se han dado a esta tarea experimental, destacan los trabajos de Zwicker, Greenwood, Scharf, Hawking y Stevens[47], cuyos resultados se muestran en la figura 2.8.

En la figura 2.8 los valores de la línea sólida fueron obtenidos por Zwicker[50] como promedios de los siguientes experimentos:

1. Empleo de un sonido complejo formado por dos tonos separados por Δf y comparación de su sonoridad con la de un tono fijo centrado en la banda Δf. Cuando Δf excede una banda crítica, la sonoridad del sonido complejo empieza a incrementarse.

2. Empleo de ruido de banda angosta en la presencia de dos tonos. Se empieza a incrementar Δf de los tonos, el umbral de la presencia del ruido se mantiene constante hasta que Δf sobrepasa una banda crítica, después el umbral cae y sigue descendiendo a medida que Δf se incrementa.

3. Dados dos tonos que varían en Δf, el oído detecta mejor los tonos en AM que en FM, siempre y cuando Δf sea menor a una banda crítica. En caso contrario, no hay diferencia en esta sensibilidad.

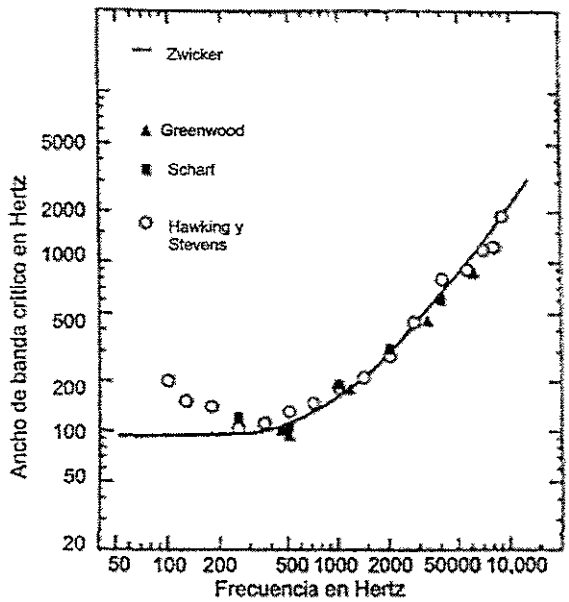


Figura 2.8. Bandas críticas en función de la frecuencia central de cada banda[47].

Los triángulos se refieren al experimento de Greenwood, donde se detecta el umbral de un tono dentro de un ruido de banda angosta. El ancho del ruido varía teniendo al tono en su frecuencia central. El umbral se modifica de acuerdo con Δf. La forma de variación es distinta dependiendo si se sobrepasa o no una banda crítica.

Los cuadrados se refieren a la combinación de dos experimentos de Scharf que son semejantes a los dos primeros de Zwicker. Los resultados en círculos se refieren a un experimento semejante al de Fletcher hecho por Hawking y Stevens, donde se calcula la banda de ruido que contribuye al enmascaramiento de un tono situado en su frecuencia central. Se obtienen estimaciones del umbral de cada banda, llamadas cocientes críticos, que son aproximaciones a las bandas críticas.

Para ruido blanco y sonidos complejos, ambos de banda angosta Δf, bajo una gran variedad de condiciones[47], se ha demostrado que la sonoridad del ruido o del sonido es independiente de Δf siempre y cuando éste no rebase el ancho de la banda crítica donde esté

situado, y es igualmente sonoro que un tono situado en la frecuencia central de la banda. Cuando Δf es mayor a la banda crítica, la sonoridad del sonido complejo o del ruido aumenta.

Banda	Frec. Central (Hz)	Ancho de la banda	Frec. Inferior (Hz)	Frec. Superior (Hz)
1	50	-	-	100
2	150	100	100	200
3	250	100	200	300
4	350	100	300	400
5	450	110	400	510
6	570	120	510	630
7	700	140	630	770
8	840	150	770	920
9	1,000	160	920	1,080
10	1,170	190	1,080	1,270
11	1,370	210	1,270	1,480
12	1,600	240	1,480	1,720
13	1,850	280	1,720	2,000
14	2,150	320	2,000	2,320
15	2,500	380	2,320	2,700
16	2,900	450	2,700	3,150
17	3,400	550	3,150	3,700
18	4,000	700	3,700	4,400

Tabla 2.1 Ejemplo de bandas críticas en el intervalo de 0-4 kHz[47]

Experimentos más recientes sobre bandas críticas han sido aplicados a sonidos. French y Steinberg[51] midieron la inteligibilidad de palabras que pasaron por filtros pasobajas y pasoaltas, encontraron que una partición en 20 bandas frecuenciales contribuye igualmente a la inteligibilidad cuando se aproximan a las bandas críticas. Morton y Carpenter[47] obtuvieron que las distribuciones de energía de dos formantes que caen en una banda crítica son más

difíciles de identificar que cuando los dos tonos está separados por más de una banda crítica. De esto, supusieron que el oído tiende a integrar las energías que caen en una banda crítica, obscureciendo las variaciones de energía en la banda.

Chaves y Sharf[47] demostraron que es más difícil detectar la diferencia de intensidades entre dos tonos dentro de una banda crítica que cuando los tonos se encuentran fuera de una banda crítica. Esto implica que los formantes son más fáciles de identificar cuando las armónicas están separadas por más de una banda crítica. Zwicker[50] construyó un dispositivo electrónico análogo al oído, que incluía un analizador de espectro en bandas críticas. Demostró que la información más relevante se mantenía cuando se realizaba este filtrado y su procesamiento. Se pudieron identificar los dígitos usando estos filtros. Los anchos de las bandas críticas corresponden aproximadamente a espaciamientos de 1.3 mm en la membrana basilar, de manera que 24 filtros pasobandas modelan adecuadamente a esta membrana. Sin embargo, no se ha demostrado que el oído filtre los sonidos en bandas críticas.

Este resultado sugiere que una descomposición en bandas críticas es una forma natural de filtrar la voz, con objeto de retener el máximo de información. Los laboratorios de la Universidad de California en Davis fueron los primeros en utilizar esta descomposición en procesamiento de voz, en especial en el reconocimiento de palabras. Las bandas críticas utilizadas en este trabajo se basan en la tabla 2.1, los valores de cada banda, en barks, se pueden aproximar por la ecuación

$$z = 13 \arctan 0.76 f + 3.5 \arctan\left(\frac{f}{7.5}\right)^2 \qquad \text{Ec 2. 3}$$

donde f está en hertz y z en barks, que obtiene resultados muy aproximados a los de la tabla 2.1, especialmente en bajas frecuencias.

Las mejores aproximaciones de bandas críticas se muestran en la tabla 2.1, con variaciones de 15% entre distintas personas. En esta tabla se puede observar que las bandas críticas son casi de ancho constante en la escalas de mels; el ancho de las bandas varía de 50 a 100 mels en bajas frecuencias, a aproximadamente 250 mels a 13 500 Hz.

En la figura 2.9 se muestra un ejemplo de decimación en bandas críticas, aplicada a una trama de 256 muestras dentro del fonema /u/. La primera gráfica muestra la trama en el

tiempo; la de la derecha, el efecto de la ventana sobre la trama; la gráfica inferior a la izquierda, la magnitud de la FFT aplicada a la ventana, resultando en 128 puntos en 4 kHz; y la inferior a la derecha, el resultado de aplicar la decimación a esta magnitud. Se promedian los valores de la FFT para obtener la amplitud de la decimación.

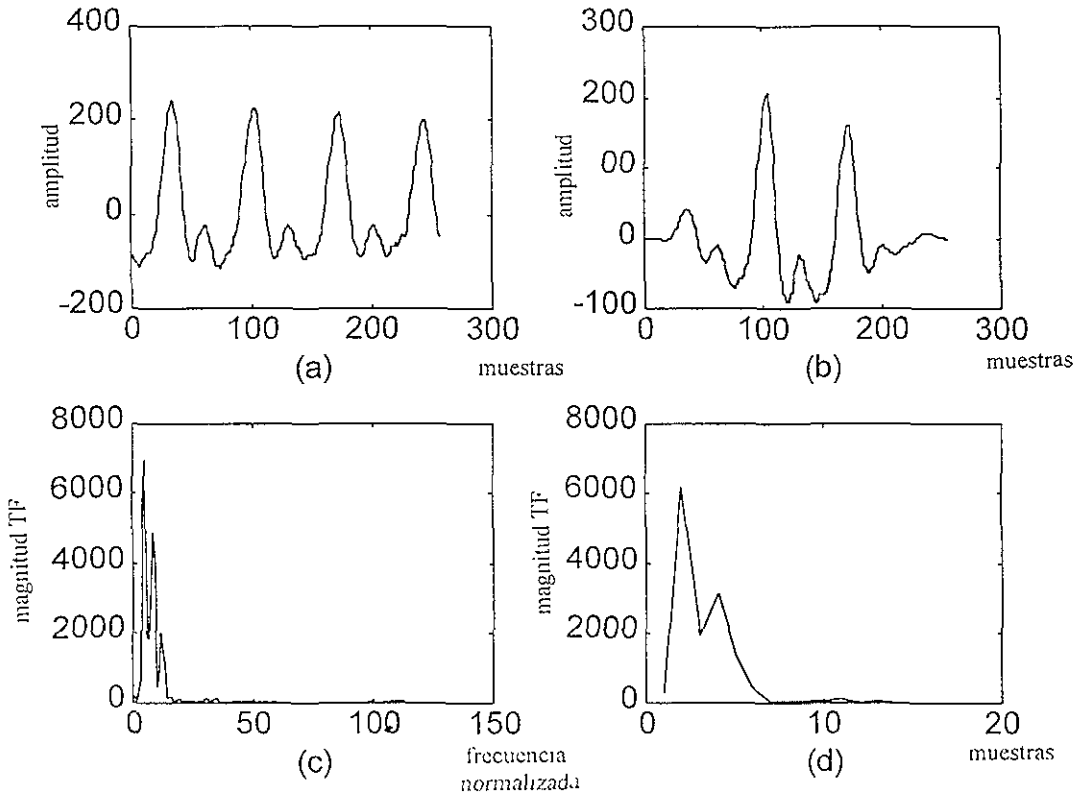
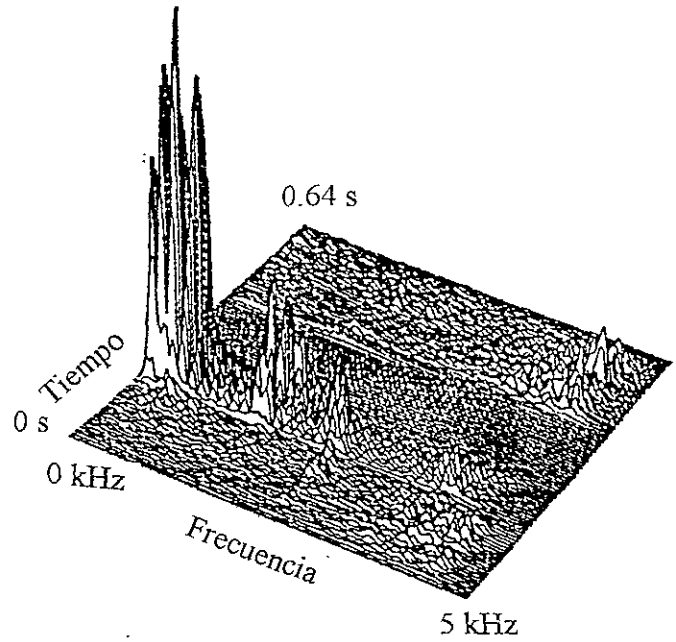


Figura 2 9. Decimación en bandas críticas aplicada a una trama de 256 muestras dentro del fonema /u/. En el inciso (a) se muestra la trama del fonema, en el (b) la trama al aplicar una ventana de Hanning, el (c) muestra la magnitud de la FFT, y el (d) la magnitud decimada en 18 bandas.

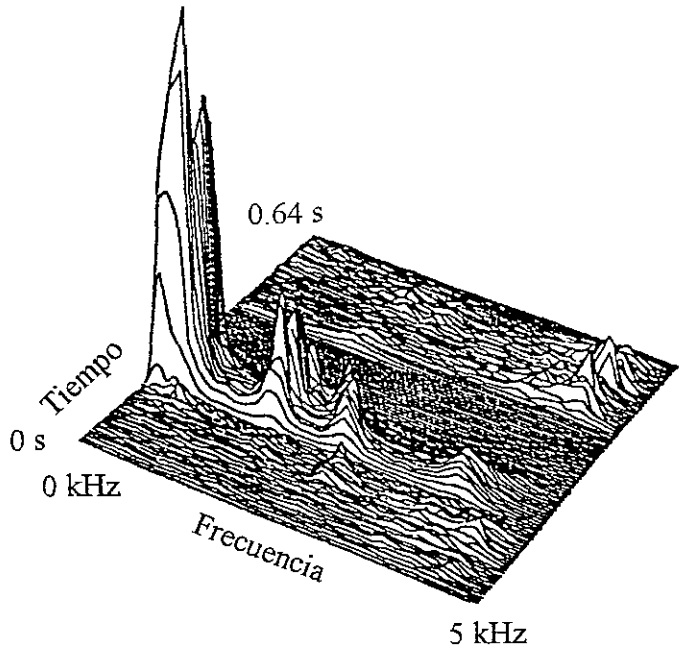
En los laboratorios de la Universidad de California en Davis, se diseñó un espectrograma tridimensional, se aplicó tanto para después del cálculo de FFT como posterior al uso de bandas críticas, para este ultimo caso se tiene que conservar el tono para reconstruir el espectrograma. El resultado se muestra en la figura 2.10 para la palabra en inglés *six*.

En la figura 2.10 puede observarse más claramente que en la figura 2.9 cómo se conserva la envolvente de la señal de voz, eliminando las pequeñas variaciones. El objetivo es eliminar las variaciones pequeñas hasta un nivel que no se pierdan las distinciones entre fonemas semejantes, lo cual se logra con esta decimación. En el caso particular de la palabra *six*, se

distingue el comportamiento de la oclusiva /k/, que en algunas repeticiones se confunde con fin de palabra.



(a)



(b)

Figura 2.10. Espectrograma tridimensional después de aplicar tramas de 128 puntos con ventanas de Hanning y (a) la FFT y, (b) bandas críticas de 18 bandas, de la palabra en inglés *six*.

Capítulo 3

SEGMENTACIÓN ACÚSTICA

En la figura 3.1 se muestra nuevamente el diagrama de bloques del preprocesamiento realizado en este trabajo. El bloque de segmentación acústica se desarrollará en este capítulo.

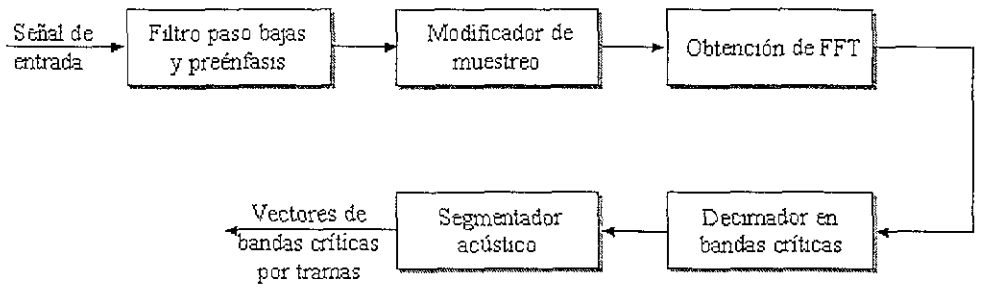


Figura 3.1 Diagrama de bloque de la fase de preprocesamiento.

Con objeto de proporcionar elementos relacionados con la fonología, se establecen primeramente algunos aspectos básicos de la fonética articulatoria y acústica del español hablado en México, además de que existe una grave ausencia de este tipo de trabajos en nuestro país.

Posteriormente, se desarrolla el método de segmentación acústica MLR que logra una segmentación pertinente para nuestros propósitos de obtener segmentos cuasiestacionarios, mayores a los de una trama. Estos segmentos o subpalabras tienen una connotación fonética pero no segmentan en fonemas, cuando estos no están presentes firmemente en la señal, el algoritmo no los separa y los une con algún fonema adyacente.

Finalmente, se plantea la segmentación MLR como una opción para la detección de palabras a métodos tradicionales como el de energía-cruces por cero. Se comparan sus resultados con una base de palabras en inglés de bajo nivel de ruido de Texas Instruments, y con una base de palabras en español con alto ruido ambiental.

3.1 Fonética acústica

Se revisarán algunos aspectos básicos de la fonética articulatoria y acústica para sonidos en el idioma español hablado en México.

Fonema	*Letras*	*Ejemplos*	*Fonema*	*Letras*	*Ejemplos*
/a/	a	cuatro		k	kilo
/e/	e	seis		q	quórum
/i/	i	cinco		qu	queso
	y	... y ...	/l/	l	lado
/o/	o	dos	/m/	m	mano
/u/	u	uno	/n/	n	nada
	w	Washington	/ñ/	ñ	niño
/b/	b	bola	/p/	p	ópera
	v	vaso	/r/	r	trozo
/c/	ch	muchacho	/r̄/	r	rosa
/d/	d	donde		rr	carro
/f/	f	café	/s/	s	mesa
/g/	g	gato		c	cebú
	gu	guerra		x	excavar
	w	wagneriano		z	zapato
	h	huevo	/t/	t	Tela
-	h	hola	/k/-/s/	x	examen
/x/	j	caja	/S/	x	Xico
	g	gesto	/y/	y	mayo
	x	México		ll	llama
/k/	c	casa			

Tabla 3.1. Fonemas del alfabeto español de México[52][53][54][55].

Este estudio es breve y se tiene la limitante de que en el país, como en muchos países, existe una gran diversidad de acentos.

El número de letras del alfabeto español quedó establecido en 1994 y reafirmado en 1999, con 27 letras , 5 vocales y 22 consonantes[56]. Como se observa en la tabla 3.1 los fonemas asociados casi igualan al número de palabras, los fonemas están expresados con los símbolos de la IPA, a excepción de /ñ/ por no tener el símbolo original.

Cabe mencionar que existen 5 dígrafos en nuestro idioma, esto es signos ortográficos compuestos de dos letras diferentes, y son: *ch*, *gu*, *ll*, *qu*, y *rr*, que se incluyen en la tabla 3.1. Las antes letras *ch* y *ll* constituyen un fonema en particular, la *gu* adquiere el fonema de *g* o bien concatena los fonemas de *g* y de *u*, asimismo el dígrafo *qu* adquiere el fonema de *q* o suma los fonemas de *q* y de *u*, para el dígrafo *rr* su fonema es igual a una de las variante múltiple del fonema de la letra *r*.

En el centro del país, es destacable la variación de los fonemas de *x* y *z* al fonema /s/ y del yeísmo de *ll*, la letra *x* también adquiere combinaciones de los fonemas /k/ y /s/, así como /g/ y /s/ aunque esta última es menos común en nuestra región del país. Si bien en los países hispano parlantes la letra *h* no tiene un fonema asociado, en la región central del país puede variar al fonema /g/.

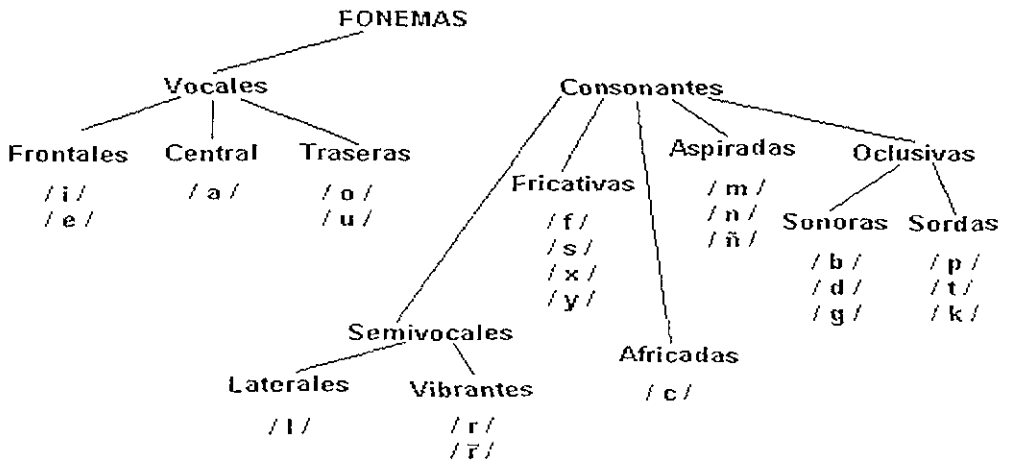


Figura 3.2. Clasificación de los símbolos fonéticos del español de México[52][53][54][55].

Los fonemas en español se pueden clasificar como se muestra en la figura 3.2, aunque estos conjuntos no son disjuntos, y por otro lado la clasificación no incluye alófonos de algunas

letras como *b,c,d* que adquieren sonidos fricativos, y tampoco incluye fonemas que se cancelan en la región central de nuestro país, pero que existen en otras regiones y más aún en otros países.

3.1.1 Vocales

Los fonemas vocálicos corresponden a las cinco vocales del alfabeto. Todos son articulaciones abiertas de corta duración, completamente sonoras, ya sea que estén acentuadas o no , sus espectros se muestran en la figura 3.3. No existen fonemas vocálicos en español con el estilo de diptongos como en el inglés.

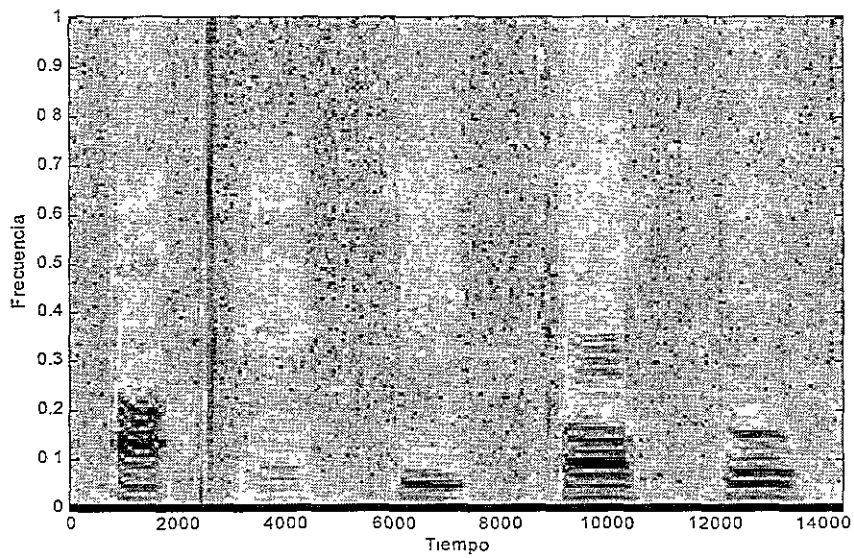


Figura 3.3. Espectogramas de las vocales del español de México.

Los tres primeros formantes vocálicos tienen en promedio las frecuencias que se muestran en la tabla 3.2. De esta tabla se puede verificar que entre mayor es la sección del tracto vocal mayor es el formante $F0$. Así también, entre más adelante y elevada esté la lengua, mayor es el formante $F1$, se tienen los casos extremos de /i/ y /u/.

Los espectros en la figura 3.3 ejemplifican los valores mostrados en la tabla 3.2 en el habla de un adulto hombre. La frecuencia está normalizada y el valor unitario corresponde a 4 kHz. Por ejemplo, en /a/ la fundamental es baja, alrededor de 600 Hz, la primer armónica está alrededor de 1 kHz, y la segunda $F2$ está en 2 kHz. Para /e/ $F1$ y $F2$ tienen baja energía y

$F0$ se sitúa alrededor de 300 Hz. Para /i/ el formante $F1$ ocupa frecuencias alrededor de 2 300 Hz, F2 no se observa, aunque existe l. Para /o/ y /u/, $F2$ está fuera del intervalo de la figura y tanto $F0$ como $F1$ ocupan frecuencias alrededor de 300 y 500 Hz. En estos ejemplos, se observa que la fundamental y armónica no se presentan única y claramente, otras armónicas resultantes de combinaciones de los fonemas aparecen constantemente e impiden una observación precisa. Sin embargo, es esta característica lo que le da riqueza a la señal y diferencia a los hablantes.

Vocal	*F0*	*F1*	*F2*
a	900	1300	2100
e	375	2200	2550
i	325	2300	3900
o	400	550	4300
u	325	425	4500

Tabla 3 2. Los tres primeros formantes de los fonemas vocálicos[53].

Los fonemas varían levemente de acuerdo con las posiciones de las vocales en las palabras y al dialecto o región de quien habla. Estas variantes se llaman *alófonos*. El principal parámetro articulatorio, la posición de la lengua, se conserva para los alófonos de una vocal. Sin embargo, otros parámetros de menor importancia pueden variar, como el redondeo de los labios o la duración del sonido.

Diptongos

Este término se refiere al monosílabo que empieza en o cerca de la posición articulatoria de una vocal y se mueve hacia la posición de otra vocal. La vocal con la mayor abertura del tracto vocal se llama núcleo silábico (vocal fuerte), la otra vocal con la menor abertura se llama sílaba marginal (vocal débil).

Un diptongo puede ser creciente o decreciente. El primero existe cuando el núcleo silábico precede al margen silábico, el segundo viceversa. Para diptongos crecientes, al margen silábico se le llama también semiconsonante[52]. Hay dos semiconsonantes crecientes [j] y [w] y ocho diptongos crecientes:

- [ja] "hacia", [je] "tiempo", [jo] "radio", [ju] "ciudad".

- [wa] "agua", [we] "suelo", [wi] "ruido", [wo] "antiguo".

Para diptongos decrecientes el margen silábico se llama también *semivocal*. Las semivocales son distintas de las semiconsonantes por su posición en el diptongo y por la forma de articulación. En las semivocales existe una articulación mayor y de más duración. Existen dos semivocales [i] y [u]y seis diptongos decrecientes:

- [ai] "aire", [ei] "seis", [oi] "hoy".

- [au] "causa", [eu] "feudo", [ou] "lo unió".

En la figura 3.4 se muestran los espectros para el diptongo creciente en "radio" y para el diptongo decreciente en "seis". Para el primero, es dominante el alto contenido en frecuencia de la vibrante /r/ y las transiciones en los formantes de las vocales /i/ y /o/. Para la palabra "seis" es importante anotar el alto contenido en frecuencia del sonido sordo /s/ y las transiciones crecientes para los fonemas vocálicos.

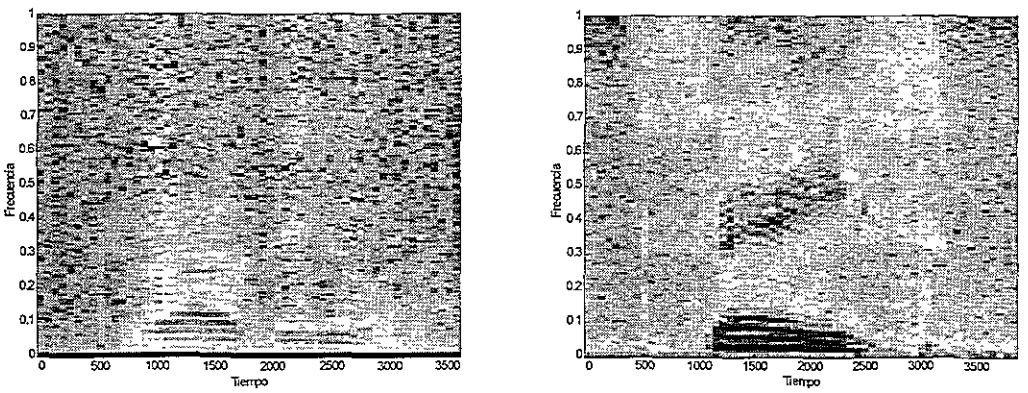


Figura 3.4. Espectros de "radio" y "seis" que muestran diptongos crecientes y decrecientes, respectivamente.

En los triptongos una vocal constituye el núcleo silábico y las otras son semiconsonantes o semivocales, dependiendo si están antes o después del núcleo. La mayoría de los triptongos corresponde a la forma verbal del pronombre "vos", no usado en México.

3.1.2 Consonantes

Las consonantes se pueden clasificar de acuerdo con las maneras de articulación descritas en la figura 1.1, es pertinente señalar que la clasificación no es única, así como tampoco el tipo de categorías. Éstas se refieren a los grados de constricción del punto de articulación y la manera en que se exhala para el siguiente sonido. Se pueden clasificar en las siguientes categorías[52][53][54]:

- *Africadas.* Existe un cierre inicial del tracto vocal seguido de una expiración gradual que produce turbulencia.

- *Fricativas.* El tracto vocal está abierto parcialmente en el punto de articulación y el velo está cerrado. Ruido se genera en el punto de articulación.

- *Laterales.* EL tracto vocal está cerrado en el punto de articulación pero abierto a los lados.

- *Nasales.* El tracto vocal está cerrado en el punto de articulación y el velo está abierto.

- *Oclusivas.* El tracto vocal está cerrado en el punto de articulación, el pasaje nasal está cerrado, y existe una exhalación limpia y cortante.

- *Semivocales.* El tracto vocal esta parcialmente abierto en el punto de articulación sin turbulencia.

- *Vibrantes.* Existe una abertura y cerradura oscilatorias en la lengua, que es el punto de articulación, seguida de una exhalación gradual que produce turbulencia.

Todos los fonemas que corresponden a las vocales, diptongos, semivocales, y nasales se conocen colectivamente como *sonoros*. Los sonidos sonoros excitan al tracto vocal con pulsos cuasi-periódicos originados por la vibración de las cuerdas vocales. En contraste, las restantes clases son excitadas fundamentalmente en punto de constricción del tracto vocal y se denominan *sordos*.

Si bien las formas de articulación dividen los fonemas en categorías muy amplias basadas en diferencias en la excitación, el lugar de articulación identifica diferencias en el tracto vocal

de acuerdo con el punto máximo de constricción en el tracto vocal y permiten diferenciar más sutilmente los fonemas que tienen la misma forma de articulación.

A pesar de que las formas de articulación son muy consistentes entre los idiomas, no ocurre así con el punto de articulación. Casi sin excepción, los idiomas emplean vocales, nasales, oclusivas y fricativas; sin embargo, el número y lugar de articulación en cada forma varían enormemente en los idiomas.

A lo largo del tracto vocal, existen aproximadamente ocho regiones o puntos de articulación que se asocian con las consonantes; sin embargo, un idioma sólo utiliza un número reducido de ellos, los más importantes son[52][53][54]:

- *Alveolar.* La punta de la lengua se acerca o toca la punta alveolar en el techo de la boca.
- *Dental.* La punta de la lengua hace contacto con la parte posterior de los dientes superiores.
- *Interdental.* Cuando la punta de la lengua toca sólo el borde de los dientes superiores.
- *Labial.* Existe una constricción en los labios.
- *Bilabial* . Denota constricción en ambos labios hasta un punto de contacto, mientras que
- *Labiodental* . Denota contacto del labio inferior con los dientes superiores.
- *Palatal.* El predorso de la lengua se eleva y toca el palador duro.
- *Velar.* El postdorso de la lengua se aproxima al paladar suave.

Para algunos casos, el punto de articulación está afectado fuertemente por el ambiente fonético de su alrededor. La descripción acústica de las consonantes es más compleja que el espectro de estado fijo que se requiere para la descripción acústica de las vocales, esto se debe a que muchos de los rasgos acústicos de las consonantes son dinámicos por naturaleza. Debido a esta naturaleza transicional de los rasgos de las consonantes, es necesario describirlos en conjunción con fonemas adyacentes. De hecho, las relaciones interfonéticas son tan complejas e interdependientes que, en muchos casos una consonante es descrita

como la parte inicial o final de una vocal. En bastantes situaciones, el rasgo distintivo de una consonante ocurre en los bordes de los segmentos adyacentes.

Oclusivas

Existen seis fonemas oclusivos: /b/, /d/, /g/, /k/, /p/ y /t/, sus diferencias se muestran en la tabla 3.3.

Fonema	*Forma de articulación*	*Lugar de articulación*	*Ejemplo*
/b/	sonoro	bilabial	"bien"
/d/	sonoro	dental	"donde"
/g/	sonoro	velar	"gloria"
/k/	sordo	velar	"casa"
/p/	sordo	bilabial	"ópera"
/t/	sordo	dental	"tela"

Tabla 3.3. Consonantes oclusivas[52][53][57].

En nuestro país construimos muchas sílabas abiertas, este hecho genera fonemas oclusivos después del núcleo silábico y modifica a las fricativas, especialmente, para fonemas sonoros, de manera tal que si bien en otros países las letras *b,d* y *g* adquieren también un fonema fricativo, éste se pierde en esta región.

Las consonantes oclusivas sonoras se dan en dos posiciones: después de una consonante nasal y una pausa para /b/ y /g/, y después de una nasal lateral y una pausa para /d/. En cualquier otra posición, las oclusivas tienden a moverse a fricativas. Por ejemplo, /b/ es oclusiva en "un bote", pero tiende a fricativa en la frase "ese bote".

En la figura 3.5 podemos observar el bajo contenido de energía de la oclusivas /p/, /t/, /d/ y /b/. En el caso de oclusivas sordas la energía es más débil y de más baja frecuencia, como por ejemplo /p/ y /t/. Los espectros de las oclusivas en español son muy similares a los de sus contrapartes en inglés. Es decir, un espectro corto precedido o seguido por un periodo largo sin energía arriba de las componentes sonoras, por ejemplo, arriba de 300 Hz, esta característica se muestra claramente en la figura 3 6 para la letra *x*.

Los tres tipos de sonidos oclusivos, de acuerdo con el punto de articulación, tienen en general, las siguientes características espectrales[53]:

- /p/ y /b/ tienen concentraciones de energía en bajas frecuencias, esto es de 500 a 1 500 Hz, donde /p/ tiene un espectro más débil que /b/. En al figura 3.5 se muestra la alta energía en bajas frecuencias del fonema /b/.

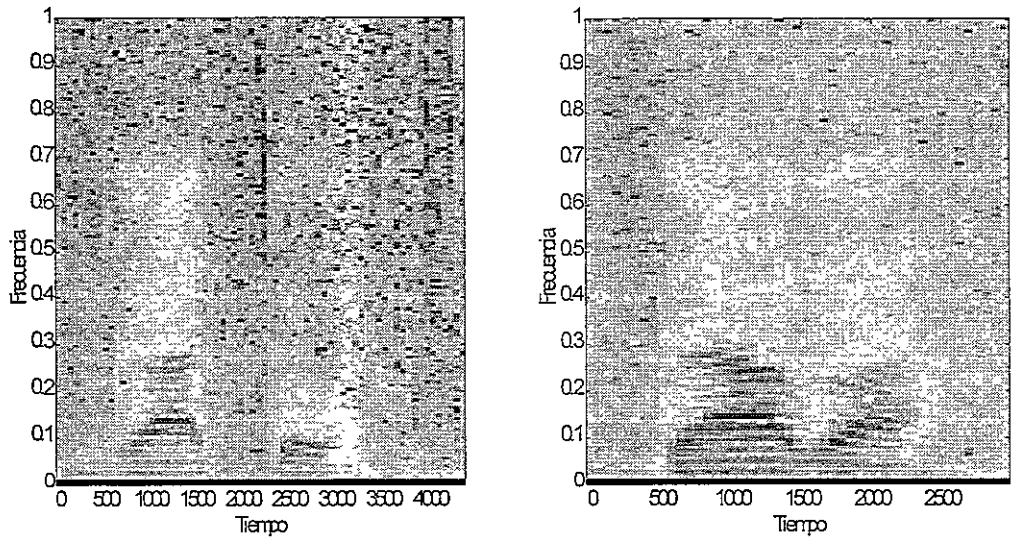


Figura 3.5. Espectros de la oclusivas "gato" y daba".

- /t/ y /d/ tienen concentraciones de energía en bajas y altas frecuencias, éstas últimas están arriba de 4 000 Hz con un espectro fuerte para esa región. Para ambos fonemas se puede observar en la figura 3.5 que su energía aunque baja se concentra en bajas frecuencias . la /d/ en 100 Hz y /t/ alrededor de 200 Hz y 500 Hz, la energía arriba de 4 kHz no es observable.

- /k/ y /g/ tienen concentraciones de energía en frecuencias medias, esto es, 1 000 a 2 000 Hz con un espectro concentrado. En la figura 3.5 se observa que para la /g/ tiene la mayor energía en 200 Hz, con otras destacables alrededor de 400 Hz y 800 Hz. En la figura 3.7 se muestra un alófono de la letra h en la oclusiva /g/, con energía destacable en bajas frecuencias.

Fricativas

En la región central del país existen cuatro fricativas: /f/, /y/, /s/ y /x/. Los fonemas pueden distinguirse por las características mostradas en la tabla 3.4.

Fonema	*Forma de articulación*	*Lugar de articulación*	*Ejemplo*
/f/	sonoro	labiodental	"café"
/y/	sordo	palatal	"cónyuge"
/s/	sordo	alveolar	"mesa"
/x/	sordo	velar	"caja"

Tabla 3.4: Consonantes fricativas[54][58].

Los sonidos fricativos tienen una forma de onda cuasi-aleatoria con suaves cambios. Las fricativas tienen un componente en frecuencia baja cercana a los 300 Hz resultante de vibración de las cuerdas vocales. Esta franja cambia abruptamente al primer formante del sonido subsecuente. Existen formantes para sonidos sordos no presentes en sonidos sonoros.

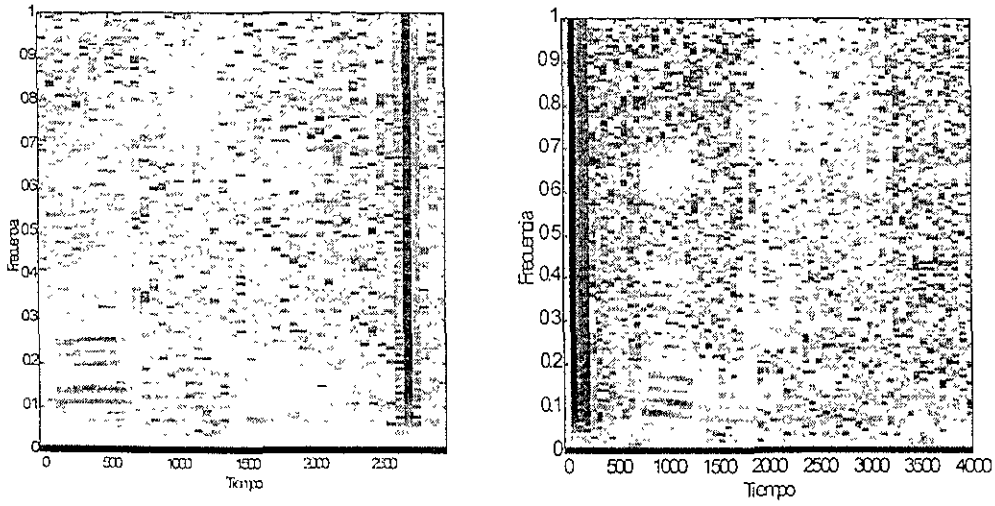


Figura 3.6. Espectros de la fricativas "Jasso" y "Fox".

Los formantes principales se encuentran en[53].

- Para /f/ y alófonos en 0-400, 1400-2200 Hz, 2900-4000 Hz y 6000-8000 Hz. En la figura 3.6 se observa que para /f/ predominan energías un poco arriba de 1 kHz.

- Para /s/ y sus alófonos en 0-500 Hz, 2600-3600 Hz y 5000-8000 Hz. En la figura 3.6 se observa que para /s/ predominan energías alrededor de 3 kHz.

- Para /y/ y sus alófonos en 0-1000, 3000-3500 Hz y 3750-5800 Hz, se caracteriza por su corta duración. En la figura 3.8 se muestra su baja energía en la frase "la llama" al encontrarse entre la vocal *a,* que es una vocal fuerte.

- Para /x/ y sus alófonos en 0-600 y 2200-3000 Hz. En este caso predomina la intensidad en las frecuencias centrales. En la figura 3.6 se observa que /x/ a pesar de ser sonora su energía se pierde casi totalmente por la vocal *a*.

Como ya se apuntó, en la figura 3.6 se pueden observar tres de estos fonemas , que muestran una característica común y muy destacable, que son sus bajas energías.

Africadas

El único sonido africado del español es /c/. Este corresponde a la letra *ch*, y es fonema sordo palatal; por ejemplo, en *muchacho*. La característica sorda, con alto contenido en altas frecuencias, se puede observar en la figura 3.7; éstas definen sus transiciones.

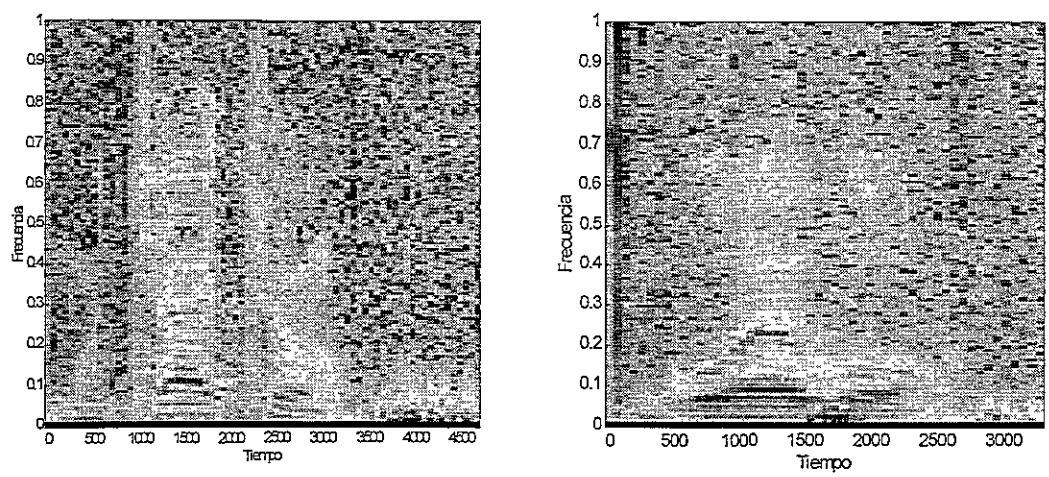


Figura 3.7. Espectro de la africada *ch* en la palabra "muchacho" y de alófono oclusivo de *h* en la palabra "huevo".

Nasales

Los tres fonemas nasales /m/, /n/ y /ñ/ son sonoros. El fonema /m/ es bilabial y corresponde a la letra *m*. El fonema /n/ es velar y corresponde a l*n*. Sin embargo, puede ser modificado en distintos alófonos.

El fonema /ñ/ es bilabial y corresponde a la letra *ñ*, no tiene alófonos. Las características de las transiciones para /m/. /n/ y /ñ/ son idénticas a aquellas que caracterizan a las oclusivas /p/, /t/ y la fricativa /x/, para las primeras cuatro formantes. Sin embargo, su energía es mayor para altas frecuencias y se extiende hasta 8 000 Hz. En la figura 3.8 el contenido en altas frecuencias no se muestra; sin embargo, se observan características semejantes a formantes vocálicos, se observa para el fonema /m/ la mayor presencia de energía en altas frecuencias para la frase "la llama" que en la palabra "muñon". Para los otros dos formantes, se observa el predominio de formantes en bajas frecuencias y la presencia de energía en frecuencias medias del espectro .

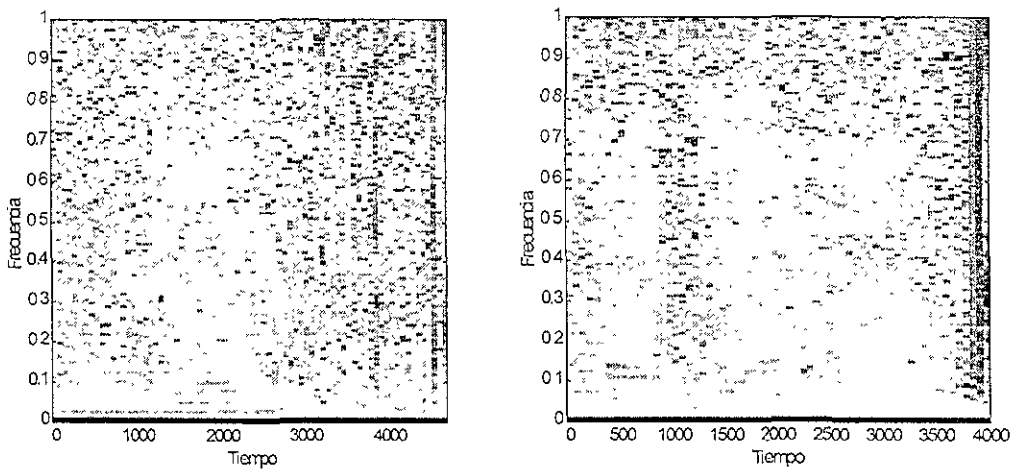


Figura 3 8. Espectros de las nasales en la palabra "muñon" y lateral en "la llama".

Semivocales

Se tiene el fonema lateral /l / que es palatal, tiene los alófonos interdentales y dentales. El primero, precede a fonemas fricativos interdentales sordos, por ejemplo en "un dulce". El segundo cuando precede a una consonante dental, por ejemplo en "el toro" También se tienen los fonemas vibrantes /r/, que corresponde a la letra "r", y /r/ que corresponde a la letra "r" o a la letra "rr".

La configuración de los tres primeros formantes de una lateral la definen de manera precisa. En la figura 3.8 se pueden observar los tres primeros formantes del fonema /l/ en la frase "la llama". Es pertinente mencionar que en nuestro país el formante de *ll* adopta el fonema /y/, como se muestra en la misma figura. En la figura 3.9 se pueden observar los formantes de mayor energía en medias frecuencias para el fonema / $\bar{r}$ / comparado con /r/.

Los espectros de las vibrantes muestran interrupciones intermitentes por intervalos de silencio. La segunda característica más importante son los valores de $F1$ a $F3$ en su inicio y en su fin. Por ejemplo, en la tabla 3.5 se muestran los primeros formantes de las palabras "lega" y "rezo".

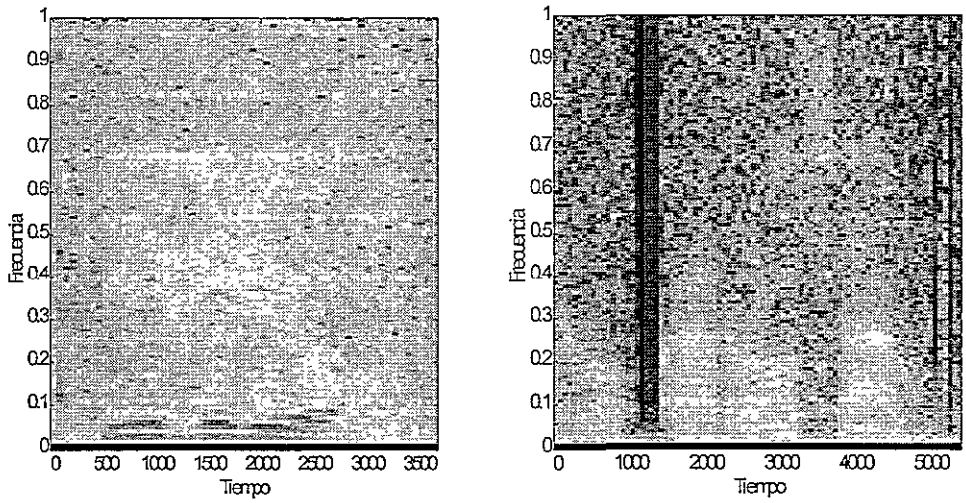


Figura 3.9. Espectros de las vibrantes en la palabras "vidrio" y "rocarosa".

Formante	*"lega"*	*"rezo"*
F0	400-475	500-550
F1	1900-2100	1750-2000
F2	2400-2500	2200-2450
F3	4500-4500	4250-4500

Tabla 3.5. Primeros formantes de las semivocales /l/ y / $\bar{r}$ /.

3.2 Segmentación acústica

Ahora se decomponen las señales de voz en segmentos de longitud variable cuasiestacionarios, llamados subpalabras acústicas, porque esta segmentación se diseña con base en cambios significativos del espectro de la señal de voz. Aparentemente no existe ninguna correspondencia con el contenido lingüístico de la señal, sin embargo, existe una fuerte relación entre ambos tipos de segmentaciones.

Se establece en señales de voz un paralelismo entre diferencias "acústicas" y diferencias en los espectros, lo cual no es fortuito ya que los cambios de fonemas se reflejan en sus propios espectros, con un comportamiento monotónico en ambos.

Las técnicas MLR y de variación de los valores propios de la KLT proporcionan segmentaciones precisas de las palabras, la última aún más que la primera. Ambas se han diseñado para ajustar el número de segmentos a un valor determinado, o bien señalar un umbral y dejar el número de segmentos libre.

3.2.1 Método MLR

El algoritmo es diseñado con base en cocientes de máxima similitud (MLR) que detectan los cambios de espectro[59][60][61]. El método se divide en dos etapas secuenciales, en la primera se obtiene la detección de inicio y fin de palabra, en la segunda la propia segmentación acústica.

En ambas etapas se utilizan ventanas que al desplazarse en el tiempo realizan comparaciones MLR. Se asume por simplicidad que los vectores de bandas críticas 18-dimensionales de cada trama, $\underline{C}(n)$, son gaussianos i.i.d. con media cero.

La formulación teórica se deriva de la expresión de la hipótesis

$$H_0 : \sum = \sum_0$$

$$H_1 : \sum \neq \sum_0 \qquad \text{Ec 3. 1}$$

con base en el segmento $[\underline{C}(1),\underline{C}(2),\ldots,\underline{C}(N)]$, las componentes de Σ ,$(\sigma_1^2,\sigma_2^2,\ldots,\sigma_J^2)$ son desconocidas y las componentes de $\underline{C}$ conocidas.

La función de similitud para el segmento asumiendo independencia entre tramas es

$$L(\underline{\mu},\Sigma\ ) = (2\pi)^{-JN/2}\,|\Sigma\ |^{-N/2}\exp\left[-\frac{1}{2}\sum_{n=1}^{N}\underline{C}(n)^T\Sigma\ ^{-1}\underline{C}(n)\right] \qquad \text{Ec 3. 2}$$

y el cociente de máxima similitud es

$$\lambda = \frac{L(\underline{0},\Sigma_0\ )}{\underset{\Sigma}{máx}\,L(\underline{0},\Sigma\ )} \qquad \text{Ec 3. 3}$$

esto es, el numerador es la función de máxima similitud para $(\underline{\mu},\Sigma\ )$ en el espacio paramétrico restringido por la hipótesis $(\underline{\mu}=\underline{0},\Sigma=\Sigma_0)$ y el denominador es el máximo sobre el espacio paramétrico ($\underline{\mu}=\underline{0},\Sigma$ positiva definida). Cuando Σ está sin restricciones, el máximo ocurre cuando está definido por el estimador de máxima similitud Σ en el cual

$$\hat{\sigma}_J^2 = \frac{1}{N}\sum_{n=1}^{N}C_{jn}^2 \qquad \text{Ec 3. 4}$$

Al insertar $\hat{\sigma}_J^2$ en (3.3) y simplificando se obtiene la prueba MLR

$$\Lambda = \ln\lambda = \frac{N}{2}\left|\sum_{j=1}^{J}\ln\frac{\hat{\sigma}_J^2}{\sigma_{0j}^2} - \sum_{J=1}^{J}\frac{\hat{\sigma}_j^2}{\sigma_{0j}^2}\right| \begin{matrix} H_0 \\ < \\ > \\ H_1 \end{matrix} T \qquad \text{Ec 3. 5}$$

donde en el paso k

$$\Lambda(k) = \frac{N}{2}\left|\sum_{j=1}^{J}\left[\ln\frac{\hat{\sigma}_j^2(k)}{\sigma_{0j}^2} - \sum_{j=1}^{J}\frac{\hat{\sigma}_j^2(k)}{\sigma_{0j}^2}\right]\right| \qquad \text{Ec 3. 6}$$

en el cual $\hat{\sigma}_j^2(k)$ representa la variancia de la muestra en la banda de frecuencia j^{th} en el paso k, $\hat{\sigma}_{0j}^2$ es la j^{th} ,variancia de Σ_0 ,y T es el umbral determinado experimentalmente.

Para la segmentación acústica, las pruebas MLR se aplican secuencialmente a cinco tramas,

$$Y(n) = [\underline{X}(n-1), \underline{X}(n), \underline{X}(n+1)] \qquad \text{Ec 3. 7}$$

donde

$$\begin{aligned} X(n-1) &= [\underline{C}(n-2), \underline{C}(n-1)] \\ X(n) &= [\underline{C}(n)] \\ X(n+1) &= [\underline{C}(n+1), \underline{C}(n+2)] \end{aligned} \qquad \text{Ec 3. 8}$$

y en la ecuación 3.6 $\sigma_{0j}^2(n)$ representa ahora la variancia de $\underline{X}(n-1)$, y $\sigma_j^2(n)$ de $\underline{X}(n+1)$.

La figura 3.10 muestra el estadístico $\Lambda(n)$ para la detección de la palabra "six". Hay dos regiones que exceden el umbral; cada región se clasifica como voz. El silencio en la mitad de la palabra está asociado con la oclusiva /k/.

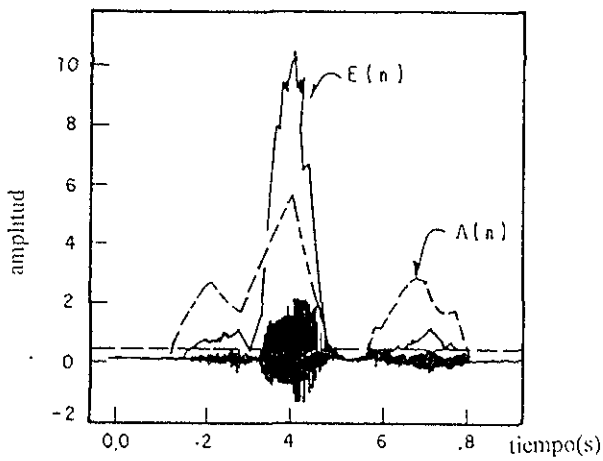


Figura 3.10 Estadístico $\Lambda(n)$ para la detección de la palabra``six".

En las gráficas 3.11, y 3.11 se muestran las señales en el tiempo y sus espectogramas, de repeticiones de las palabras *tres* y *cinco*, señalando también las segmentaciones con líneas verticales.

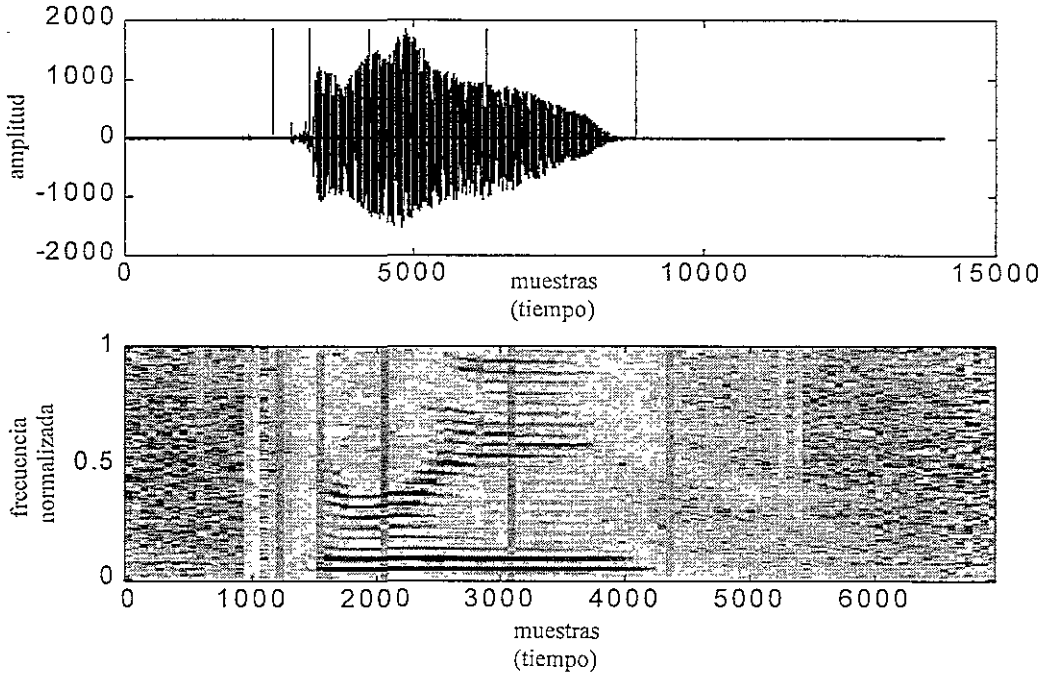


Figura 3.11. (a) Gráfica en el tiempo de la palabra "tres" y, (b) segmentación acústica de esta palabra usando MLR en 4 segmentos.

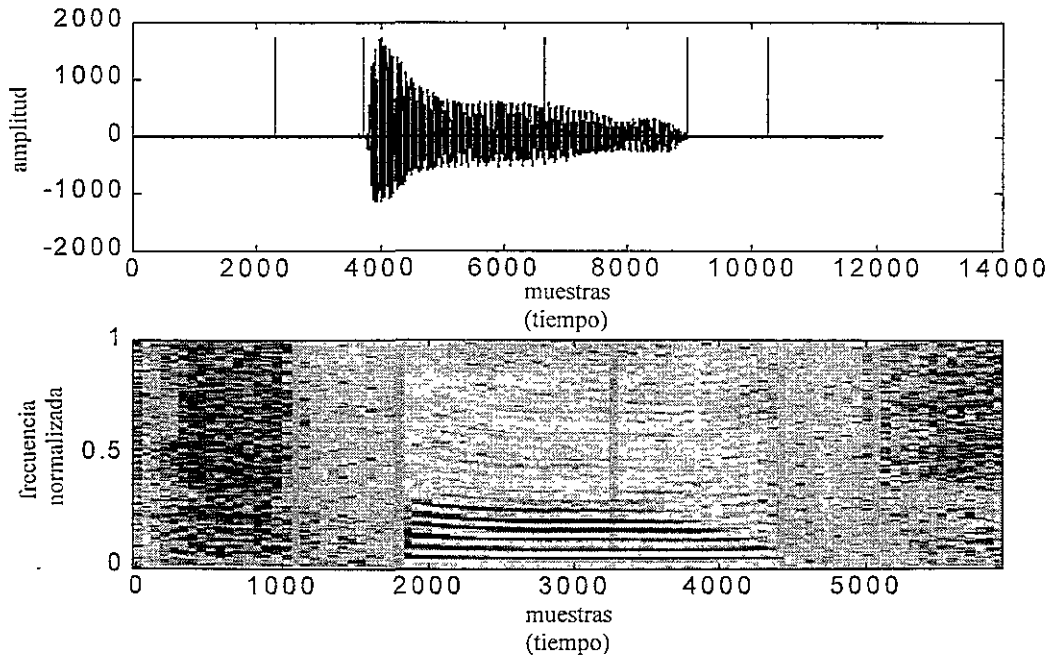


Figura 3.12. (a) Gráfica en el tiempo de la palabra "cinco" y, (b) segmentación acústica de esta palabra usando MLR en 4 segmentos .

En la gráfica 3.11 se observa la señal con cuatro segmentaciones correspondiendo cada segmento a un fonema, en la primer subpalabra está la fricativa sorda /t/ con muy baja energía, el cambio espectral entre los fonemas /r/ y /e/ se muestra en la segunda segmentación, es de particular interés observar el fonema /s/ ya sonoro en esta repetición.

En la gráfica 3.12 se observa la palabra *cinco* con cuatro segmentos, la primera subpalabra corresponde a la fricativa /s/, la segunda engloba a los fonema /i/ y /n/ y la tercera a /k/ y /o/, la última es en realidad ruido al término de la /o/, muy común en esta vocal.

3.2.2 Método de valores propios de la KLT

El método de variación de valores propios de la KLT consiste en la medición de valores propios residuales de la KLT por trama, estos valores son los 11 últimos que no han sido considerados para representar la señal con la KLT, esto es,

$$E_r = \sum_{i=8}^{18} \lambda_i \qquad \text{Ec 3. 9}$$

cuando esta energía residual sobrepasa un umbral predeterminado, se considera que termina un segmento.

Dado que se analiza trama por trama, este método es de mayor precisión que MLR, sin embargo, esta mayor precisión no modifica la precisión de reconocimiento, y en ocasiones es difícil de detectar en señales en el tiempo o en espectogramas.

En la figura 3.13, se muestra la palabra *cero* con siete, seis, cinco y cuatro segmentos. De éstas, puede observarse a través del fonema /s/ que los fonemas sordos presentan cambios espectrales fuertes que producen varias segmentaciones, sobretodo cuando se encuentran al inicio o fin de la palabra.

Asimismo, se observa que el fonema /o/ al final de la palabra tiene asociados dos segmentos en los tres primeros espectogramas, en el último segmento es donde se presenta respiración más que sonido.

El exceso de segmentos en los casos mencionados se presentan en otras palabras, y representan un factor importante por considerar en el diseño de los métodos de clasificación.

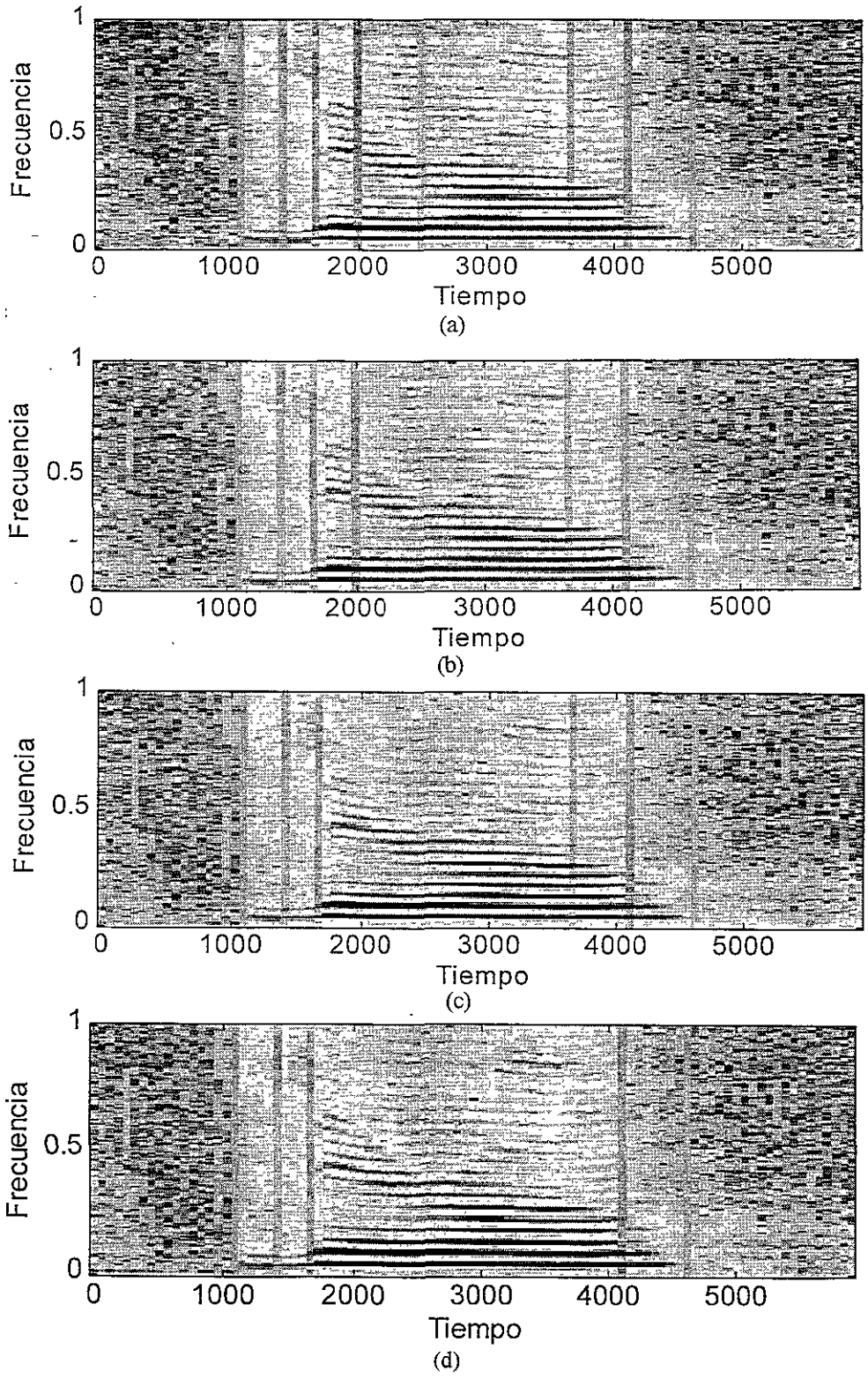


Figura 3.13. Segmentación acústica de la palabra "siete" usando valores propios de la KLT, con (a) 7 segmentos, (b) 6 segmentos, (c) 5 segmentos y (d) 4 segmentos.

3.3 Determinación de inicio y fin de palabra

En los inicios de las investigaciones sobre reconocimiento de voz, aun de palabras aisladas, se observó la ventaja de detectar inicio y fin de palabra. Obviamente, la tasas de reconocimiento mejoraban al aplicar esta detección a un archivo de voz independientemente de las técnicas de clasificación utilizadas. Sin embargo, este problema básico no ha sido resuelto en su totalidad aun para palabras aisladas, especialmente donde existe un alto nivel de ruido, lo que constata la intratabilidad de este tipo de procesamiento.

En el desarrollo de investigaciones en detección de inicio y fin de palabra, se han clasificado los detectores en explícitos, implícitos e híbridos. Los explícitos realizan la detección independientemente del procesamiento que se realice posteriormente a las palabras; los implícitos realizan primeramente cierto tipo de procesamiento, generalmente el reconocimiento de palabras, para con este procesamiento obtener inicio y fin; y los híbridos realizan una primera detección de palabras, posteriormente realizan un primer procesamiento, después ajustan los inicios y fines de palabras auxiliados por el procesamiento, y finalmente se rehace el procesamiento. En todos los casos destacan las técnicas diseñadas por Rabiner con otros autores y son universalmente utilizadas. Una primera de tipo explícito expuesta con Sambur en 1975[62], está basada en una combinación de energía corta y cruces por cero. Las del tipo implícito e híbrido las diseñó con Lamel, Rosenber y Wilpon, ambas técnicas fueron expuestas en 1980[63], donde utilizaron una etapa de clasificación DTW para mejorar la detección.

El algoritmo Rabiner-Sambur es el más utilizado. Aprovecha dos técnicas básicas del dominio del tiempo: energía y cruces por cero. Para señales de bajo ruido ambiental, la energía de la voz es mucho mayor a la del ruido ambiental y por otro lado, la densidad espectral y en consecuencia, el número de cruces por cero de los fonemas es distinto a los del ruido, de tal manera que estos dos parámetros se combinan en el algoritmo.

En principio, la idea es localizar un inicio y fin de palabra usando la energía de la señal. Para señales de voz, este método de energía sería insuficiente ya que muchas palabras inician o terminan en fonemas sordos cerrados, que se caracterizan por su baja energía. Por lo que posteriormente, se precisan estos puntos de inicio y fin, usando los cambios en cruces por cero alrededor de los puntos anteriores. Obviamente, algunos problemas aparecen cuando el

espectro de algunos fonemas sordos y cerrados se asemeja al del ruido ambiental. Otro aspecto que se considera, es el evitar la detección incorrecta de terminación de palabras cuando se tienen largos silencios en fonemas oclusivos.

Se hará una breve descripción de este algoritmo, mostrado en la figura 3.13.

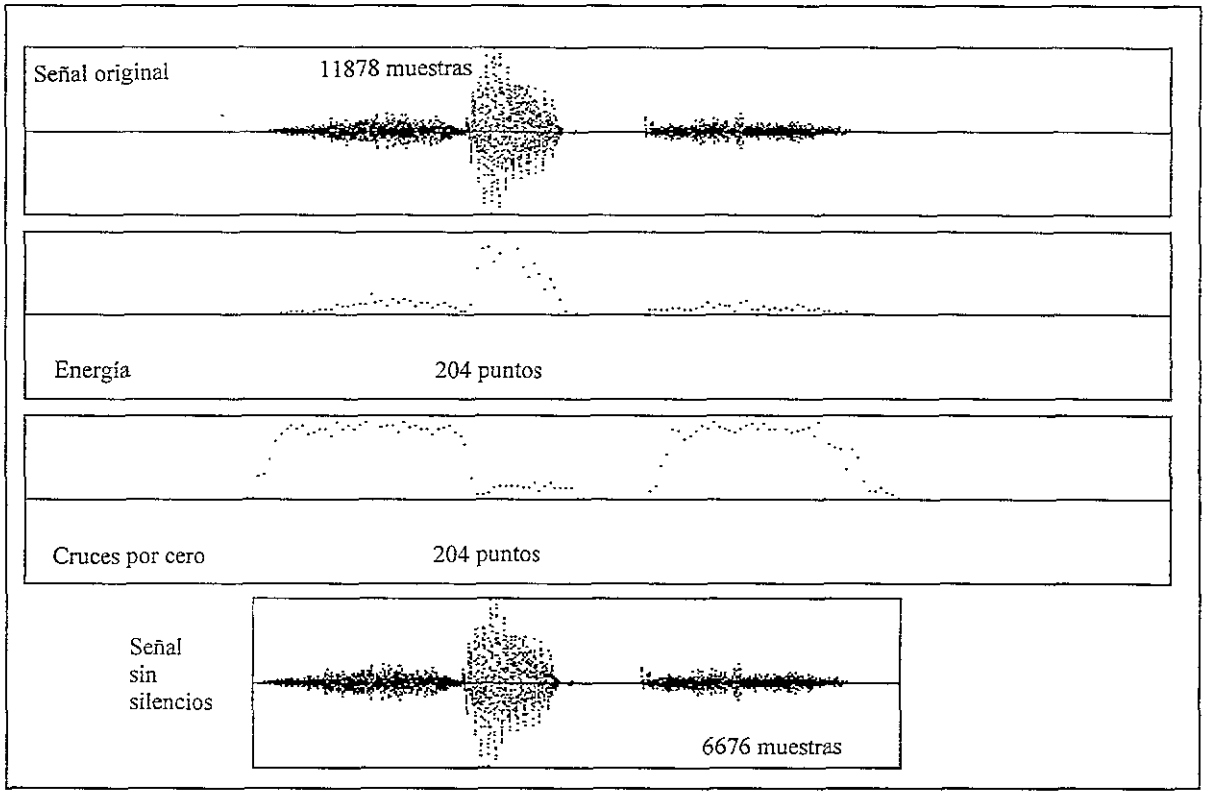


Figura 3.13. Método combinado de energía y cruces por cero.

Se calcula el número de cruces por cada trama de 10 ms y la magnitud promedio en tramas de la misma longitud. No se consideran traslapes en las tramas y se aplican ventanas de Hanning, $\omega(i,N)$, de acuerdo con las relaciones

$$E(n)=\sum_{i=-55}^{55}|s(n+i)| \qquad \text{Ec 3. 10}$$

$$ZCR(n)=\sum_{i=0}^{N-1}\frac{1}{2}|\operatorname{sgn}[x(n+i)]-\operatorname{sgn}[x(n+i+1)]|\,\omega(i,N) \qquad \text{Ec 3. 11}$$

Se considera que en los primeros 100 ms de una palabra grabada sólo existe silencio. Se calcula durante este intervalo la tasa de cruces por cero, y para ésta su media (η_{cc_1}), su desviación estándar (τ_{cc_1}) y la tasa mínima (T_{cc_1}) de 25 cruces por cero, que es un valor sobrepasado fácilmente por cualquier sonido.

Con estos elementos se establece un umbral de cruces por cero (U_{cc_1}) como

$$U_{cc_1} = \min(T_{cc_1}, \eta_{cc_1} + 2\tau_{cc_1}), \qquad \text{Ec 3. 12}$$

éste es un umbral para distinguir el silencio de la palabra en un ambiente de bajo ruido.

Con los picos de energía de la palabra ME_p y del silencio ME_s se establecen los umbrales alto (U_A) y bajo (U_B) de energía de la palabra como

$$\begin{aligned} U_A &= \min(0.03(ME_p - ME_1) + ME_s, 4ME_1 \\ U_B &= 5U_A \end{aligned} \qquad \text{Ec 3. 13}$$

donde se observa que el umbral bajo de energía es el mínimo entre cuatro veces la máxima energía del silencio y 3% de la máxima energía de la señal ajustada a la energía del silencio. El umbral alto de energía es simplemente cinco veces el umbral bajo. El índice bajo es un valor muy conservador para la energía de sonidos sordos y el umbral alto es sobrepasado fácilmente por sonidos sonoros.

El algoritmo inicia buscando el umbral bajo de energía, el primer punto que lo alcance se considera inicio de palabra. El inicio se mantiene si la energía no cae abajo de este umbral antes de que sobrepase por primera vez el umbral alto. Si no es así, el inicio se recorre al punto que sobrepase a U_B y después a U_A. Un proceso similar se usa para el fin de palabra. Con esto se busca evitar falsos inicios o fines de palabra causados por respiraciones fuertes, ruidos de micrófono, silencios de la palabra, etc., cuya energía es comparable a la de sonidos sordos.

Para eliminar los errores debidos a los distintos tipos de ruido con energía alta, el algoritmo utiliza ahora los cruces por cero

Los puntos preliminares de inicio y fin se denotan $I1$ e $F1$. Se forman los intervalos $(I1-25, I1)$ e $(F1, F1+25)$. Si en el inicio, por ejemplo, se excede este intervalo en tres ocasiones el umbral de cruces por cero U_{ccv}, el punto de inicio retrocede al primer punto que sobrepasa el umbral, llamado $I2$, en caso contrario se mantiene en $I1$, se realiza algo semejante para fin de palabra.

El análisis de cruces por cero a pesar de ser un análisis frecuencial burdo proporciona en general buenos resultados, en virtud de que el comportamiento frecuencial de gran parte de tipos de ruido es blanco o rosa, y en contraparte el de fonemas sordos débiles es violeta.

La precisión de este algoritmo es muy alta y de hecho es el más usado.

Las precisiones obtenidas por el método implícito del propio Rabiner y Lamel no aumentaron significativamente[63], en cambio si mejoraron con una técnica híbrida que no ha sido usada por dos razones, (1) en el diseño de algoritmos de reconocimiento, la detección de inicio y fin es un algoritmo previo al reconocimiento y no viceversa, es decir, si ya se obtuvo un reconocimiento de palabras pierde sentido volver a detectar inicio y fin, ya que éste lo utilizamos para incrementar la tasa de precisión en el reconocimiento, y no es pertinente realizar una interacción detección-reconocimiento por el alto tiempo de procesamiento, (2) se busca un método de detección de inicio y fin sencillo y rápido, y tanto el método implícito como el híbrido son muy lentos y con un procesamiento demasiado intenso.

Si bien los métodos implícito e híbrido no son útiles para este trabajo, Rabiner y Lamel[63] proponen un nuevo método para obtener la energía, usado para lograr una mayor separación entre señal y ruido ambiental. Sin embargo, como se demostró[64], no se mejora el rendimiento del algoritmo original de energía de Rabiner

En el método MLR, primeramente la señal de entrada se diferencia en voz y silencio. Las pruebas MLR se aplican de manera secuencial a cada tres segmentos $\mathrm{Y}(n)=[\underline{C}(n-1), \underline{C}(n), \underline{C}(n+1)]$. En la ecuación 3.6, $\hat{\sigma}_{0j}^2$ representa ahora la variancia del ruido correspondiente a la j-ésima banda de frecuencia y se obtiene en los primeros 60 ms de la señal de entrada, la cual se considera silencio. Además,

$$\hat{\sigma}_j^2(n)=\frac{1}{3}\sum_{l=n-1}^{n+1} C_{jl}^2 \qquad \text{Ec 3. 14}$$

un umbral de 0.25 se usó para esta segmentación.

Para medir la precisión de los métodos de detección, se comparan los puntos de inicio y fin obtenidos en los métodos contra una detección de referencia, ésta última se obtiene fijando manualmente los puntos al escuchar donde inicia y termina la palabra dentro de la grabación. Este es un proceso penoso ya que cada repetición debe escucharse repetidamente.

Intervalo	*Desempeño*
diferencia ≤ 128 muestras	Muy bueno
128 muestras < diferencia ≤ 512 muestras	Bueno
512 muestras < diferencia ≤ 1024 muestras	Malo
diferencia > 1024 muestras	Muy malo

Tabla 3.6. Intervalos de desempeño en la detección de palabras.

La detección se clasifica en muy buena, buena, mala o muy mala de acuerdo con la separación de los puntos de inicio y fin de los métodos con los obtenidos en la referencia. Debido a que la duración de los fonemas es muy variable, se consideran intervalos muy estrictos, de menos de un fonema sordo corto, para considerar un muy buen desempeño, y de menos de un fonema de un sonido sordo de regular duración, para clasificarlo como un desempeño bueno. Así también, se considera un desempeño malo cuando es menos de un fonema sonoro, y muy malo cuando la distancia supera más de un fonema cualquiera. Los intervalos se muestran en la tabla 3.6.

En la base de datos de Texas Instrumentos para palabras aisladas, se utilizó un subconjunto formado por 400 dígitos, 4 repeticiones de cada dígito por cada uno de 10 hablantes, para comparar los métodos MLR y Rabiner-Sambur.

Los resultados arrojan un desempeño superior, en general, del método MLR bastante señalado en el inicio de palabras y ligeramente inferior en fin de palabras, como se muestra en la tabla 3.7.

Para la base de datos en español grabada en ruido ambiente, se utilizaron 1,600 repeticiones de los dígitos, 16 repeticiones de cada dígito por cada uno de los 10 hablantees.

Los desempeños muestran ya una diferencia más significativa a favor del método MLR para inicio de palabra y se repite el comportamiento favorable en fin de palabra, como puede observarse en la tabla 3.8.

Método	*Desempeño*				
Inicio	*Muy bueno*	*Bueno*	*Malo*	*Muy malo*	*Rechazados*
Rabiner-Sambur	296	66	29	7	2
MLR	334	52	12	2	0
Fin	*Muy bueno*	*Bueno*	*Malo*	*Muy malo*	*Rechazados*
Rabiner-Sambur	338	2	32	6	2
MLR	321	42	34	1	0

Tabla 3.7. Desempeño de los métodos Rabiner-Sambur y MLR de detección de palabras, para la base de TI.

Método	*Desempeño*				
Inicio	*Muy bueno*	*Bueno*	*Malo*	*Muy malo*	*Rechazados*
Rabiner-Sambur	874	61	563	100	2
MLR	936	155	411	97	1
Fin	*Muy bueno*	*Bueno*	*Malo*	*Muy malo*	*Rechazados*
Rabiner-Sambur	977	153	271	197	2
MLR	788	217	427	167	1

Tabla 3.8. Desempeño de los métodos Rabiner-Sambur y MLR de detección de palabras, para la base de datos en español.

La combinación de ambas técnicas ofrece un mejor resultado para inicio de palabra y menos casos de mal desempeño, pero baja un poco el de muy alto rendimiento[64].

Capítulo 4.

ANÁLISIS DE VOZ POR LA KLT

El análisis de voz se ha realizado en el llamado tiempo corto, a estas ventanas se les llamó tramas, en éstas se obtuvieron distintos parámetros entre los que han destacado los coeficientes LPC y cepstral. En el reconocimiento de voz, las tramas constituyeron las unidades básicas para palabras aisladas y conectadas.

Para el reconocimiento de palabras continuas, se observó que con el uso de HMM la unidad básica pertinente era el fonema, esta unidad también se utilizó para los reconocimientos antes mencionados con HMM. Sin embargo, experimentos recientes con esta técnica han concluido que otras unidades lingüísticas de mayor duración pueden operar con mayor rapidez y con precisión semejante a la de fonemas. Esto ha llevado a una experimentación de diversas unidades básicas.

En este trabajo se presenta la utilización de una unidad para la cual se tienen espectros que representen segmentos cuasi estacionarios, esto es, en espectros planos o con variaciones lineales.

En cualquier tipo de segmentación, lineal o lingüística, los parámetros estándar, obicuos, han sido los coeficientes LPC o cepstral. Otras técnicas, desde diversos tipos de transformadas de Fourier o coeficientes residuales, no ofrecen mayores niveles de compresión y la robustez de los parámetros antes mencionados. No importando la técnica de clasificación o el tipo de reconocimiento, estos parámetros se calculan por trama.

En este trabajo se desarrolla una nueva técnica de parametrización basada en la transformada de Karhunen-Loève, cuyos parámetros se obtienen en niveles de segmentos o subplabras acústicas, y no en nivel de tramas. Como ya se mencionó en el capítulo anterior, las subplabras acústicas tienen una duración de entre 20 y 50 tramas.

Los distintos tipos de parametrizaciones usadas son insuficientes para caracterizar una subpalabra acústica. Como se tratará de argumentar, la KL tiene propiedades para caracterizar la estructura espectral de la subpalabra, más aún, si se usa su propiedad de transformada óptima se comprimirá su representación en el segmento, para utilizar una fracción muy pequeña para caracterizar los segmentos acústicos.

El argumento más fuerte contra la utilización de la KLT reside en que no es óptima, sin embargo, aprovechando las propiedades cuando se trabaja con segmentos cuasi estacionarios se obtendrá un método rápido en su obtención, y al calcular en nivel de segmentos y no tramas, se reduce el número de operaciones por realizar en la clasificación.

El análisis de transformadas ortogonales se realiza en espacios de Hilbert. Sin embargo, dado que el espacio que se maneja en procesamiento de voz es real y dimensionalmente finito es suficiente con manejar espacios con producto interno y álgebra matricial real.

4.1. La transformada de Karhunen-Loève

La trasformada de Karhunen-Loève como otras de las transformadas más conocidas pertenecen al tipo de las ortogonales, o unitarias en el caso complejo. Como es óptima, en el sentido MSE, para la aproximación por un subconjunto de su representación, es usada como referencia por otras transformadas subóptimas, que en contraparte tienen algoritmos rápidos para su obtención[65].

Sea x un vector aleatorio de dimensión N, sobre el campo $\Re$. Siempre es posible expresarlo como la suma de N vectores ϕ_i linealmente independiente en la forma

$$x = \sum_{i}^{N} \phi_i y_i = \Phi y \qquad \text{Ec 4. 1}$$

donde $\Phi = [\phi_1 \ldots \phi_N]$'es una matriz no singular, y $y = [y_1 \ldots y_N]'$. Los vectores ϕ_i conforman una base de $\Re^N$, es deseable que esta base sea ortonormal.

En ingeniería es muy importante el conjunto de transformaciones ortogonales $y = Tx$, del vector aleatorio x, ya que preservan la energía y la distancia, entre otras importantes propiedades.

Además, entre otros objetivos, es deseable comprimir la representación a $M, M \leq N$ componentes de y, y así estimar a x, entonces

$$\tilde{x} = \sum_{i=1}^{M} \phi_i y_i + \sum_{i=M+1}^{N} \phi_i b_i \qquad \text{Ec 4. 2}$$

donde b_i son los últimos coeficientes. El error introducido es

$$\varepsilon = x - \tilde{x} = \sum_{i=M+1}^{N} \phi_i (y_i - b_i) \qquad \text{Ec 4. 3}$$

o bien, el error cuadrático medio es $\varepsilon_{mse}(M) = E\{\|\varepsilon\|^2\}$. Al sustituir el vector x por su valor en la ecuación 4.1, y aplicando la condición de ortonormalidad, se tiene que

$$\varepsilon_{mse}(M) = E\{\sum_{i=M+1}^{N}\sum_{j=M+1}^{N}\phi'_i\,\phi_j(y_i - b_i)(y_j - b_j)\} = E\{\sum_{i=M+1}^{N}E\{(y_i - b_i)^2\}\} \qquad \text{Ec 4. 4}$$

Para obtener el valor óptimo de b_i, usamos el criterio de Newton,

$$\frac{\delta}{\delta b_i}E\{(y_i - b_i)^2\} = -2[E\{y_i\} - b_i] = 0 \qquad \text{Ec 4. 5}$$

el resultado es $b_i = E\{y_i\}$.

Con este valor, se puede escribir $b_i = \phi'_i\,E\{x\}$, sustituyendo está fórmula en el error MSE, se tiene,

$$\varepsilon_{mse}(M) = \sum_{i=M+1}^{N}\phi'_i\,E\{(x-\bar{x})(x-\bar{x})'\}\phi_i \qquad \text{Ec 4. 6}$$

denotando C_{xx} a la matriz de covariancias, se tiene que

$$\varepsilon_{mse}(M) = \sum_{i=M+1}^{N}\phi'_i\,C_{xx}\phi_i \qquad \text{Ec 4. 7}$$

Usando ahora el Método de Lagrange, se minimiza

$$\hat{\varepsilon}_{mse}(M) = \varepsilon_{mse}(M) - \sum_{i=M+1}^{N}\beta_i[\phi'_i\,\phi_i - 1] = \sum_{i=M+1}^{N}\{\phi'_i\,C_{xx}\phi_i - \beta_i[\phi'_i\,\phi_i - 1]\} \qquad \text{Ec 4. 8}$$

con respecto a ϕ_i, donde β_i son los multiplicadores.

De aquí, se puede verificar que $\nabla_{\phi_i}[\phi'_i\,C_{xx}\phi_i] = 2C_{xx}\phi_i = 2\phi_i$. Sustituyendo en la ecuación 4.8 se tiene que,

$$\nabla_{\phi_i}[\hat{\epsilon}_{mse}] = 2C_{xx}\phi_i - 2\beta_i\phi_i = 0 \qquad \text{Ec 4. 9}$$

de donde se obtiene que

$$C_{xx}\phi_i = \beta_i\phi_i \qquad \text{Ec 4. 10}$$

Esta última ecuación establece que la matriz Φ está formada por vectores ortonormales propios de la matriz de covariancias C_{xx}, y los multiplicadores β_i son los valores propios correspondientes. Al sustituir la expresión 4.10 en la ecuación 4.7 se obtiene que

$$\varepsilon_{min}(M) = \sum_{i=M+1}^{N} \lambda_i \qquad \text{Ec 4. 11}$$

los valores propios de estos vectores aleatorios son siempre positivos.

La combinación de x en la ecuación 4.1 se denomina expansión de Karhunen-Loève, y la transformación $y = Tx$ con los vectores propios de la matriz C_{xx} es llamada transformada de Karhunen-Loève. La minimización de $\varepsilon_{mse}(M)$ es llamada análisis por componentes principales. De los resultados se concluye que:

- La KLT es una transformada óptima en el sentido de minimizar el MSE.
- Las matrices de covariancias de los vectores y y x son similares, donde C_{yy} es una matriz diagonal formada por $C_{yy} = diag(\lambda_1, \lambda_2, \ldots, \lambda_N)$,y $C_{yy} = TC_{xx}T'$.
- Los valores propios representan las variancias de los vectores y_i.
- Los vectores propios ϕ_i no son correlacionados.

La ecuación 4.11 sugiere que en una compresión de la expansión de KL se omitan los valores propios de menor valor para obtener el menor error medio cuadrático.

La KLT es óptima también para otros criterios, entre los que destacan: la representación de mínima entropía, coeficientes no correlacionados, máximas variancias de los coeficientes y mínima representación reducida[66][67], los cuales pueden establecerse ahora como propiedades de la KLT.

El siguiente teorema muestra una conclusión semejante a la de minimización del MSE.

Teorema 1. Mínima representación reducida KLT. Sean $y = Ax$, $w = I_m y$ y $z = Bw$, entonces el error medio cuadrático, entre las secuencias x y z está dado por

$$J_m = \frac{1}{N} E\left(\sum_{n=0}^{N-1} |x(n) - z(n)|^2\right) = \frac{1}{N} Tr[E\{(x-z)(x-z)^T\}] \qquad \text{Ec 4. 12}$$

llamado error de la representación reducida, éste error es mínimo cuando, $A = \Phi$, $B = \Phi^T$, $AB = I$, y los vectores propios de Φ se arreglan de manera decreciente en correspondencia a los valores propios.

Demostración. La error medio cuadrático se puede reescribir como

$$J_m = \frac{1}{N} Tr[(I - BI_m A)C(I - BI_m A)^T] \qquad \text{Ec 4. 13}$$

para obtener el mínimo se deriva esta expresión y se iguala a cero, se obtiene $I_m B^T (I - BI_m A)C = 0$, por lo que

$$J_m = \frac{1}{N} Tr[(I - BI_m A)C] \qquad \text{Ec 4. 14}$$

$$I_m B^T = I_m B^T B I_m A \qquad \text{Ec 4. 15}$$

Para $m = N$, el valor mínimo de J_m debe ser cero, lo que requiere que $B = A^{-1}$, entonces

$$I_m B^T B = I_m B^T B I_m, \quad 1 \le m \le N \qquad \text{Ec 4. 16}$$

La ecuación 4.16 implica que $B^T B$ es diagonal. Sin pérdida de generalidad, se puede normalizar a que B sea ortogonal, entonces A también lo es y,

$$J_m = \frac{1}{N} Tr[(I - A^T I_m A)C] = \frac{1}{N} Tr[I - I_m A C A^T] \qquad \text{Ec 4. 17}$$

como C es fijo, J_m se minimiza si la cantidad

$$\tilde{J}_m = Tr[I_m A C A^T] = \sum_{k=0}^{m-1} a_k^T C a_k \qquad \text{Ec 4. 18}$$

se maximiza, donde a_k es el k-ésimo renglón de A. Para maximizar a $\tilde{J}_m$ se forma el lagrangiano, que es semejante al obtenido para MSE,

$$\tilde{J}_m = \sum_{k=0}^{m-1} a_k{}^T C a_k + \sum_{k=0}^{m-1} \lambda_k (1 - a_k^T a_k) \qquad \text{Ec 4. 19}$$

de donde al diferenciar, se obtiene la condición $Ca_k = \lambda_k a_k$, por lo que

$$\tilde{J}_m = \sum_{k=0}^{m-1} \lambda_k \qquad \text{Ec 4. 20}$$

que se maximiza con los mayores valores propios.◊

Teorema 2. Coeficientes no correlacionados. Los coeficientes de la transformada KL $\{y_k, k = 0,1,\ldots,N-1\}$ no son correlacionados y tienen media cero, esto es:

$$E[y_k] = 0 \qquad \text{Ec 4. 21}$$

$$E[y_k y_l] = \lambda_k \delta(k - l) \qquad \text{Ec 4. 22}$$

Demostración De la definición de coeficientes, se tiene que

$$E[yy^T] = [\Phi]^T E[xx^T][\Phi]^T[C][\Phi] = [\Lambda] \qquad \text{Ec 4. 23}$$

que demuestra la ec. 4.23 del teorema, la ec. 4.22 es directa.◊

Teorema 3. Distribución de variancias. Entre todas las transformaciones unitarias $y = Tx$, la transformada $KL[\Phi]^T$ produce el máximo promedio de energía en $m \le N$ muestras de y .

Demostración. Se definen $\sigma_k^2 = E[|y(k)|^2]$ y $S_m(T) = \sum_{k=0}^{m-1} \sigma_k^2$. Se observa que

$$S_m(T) = \sum_{k=0}^{m-1} (TRT^*)_{k,k} = tr(I_m T^* RT) = \tilde{J}_m \qquad \text{Ec 4. 24}$$

que es máxima cuando cuando T es la KLT. Como $\sigma_k^2 = \lambda_k$ cuando $T = \Phi^{*T}$ entonces

$$\sum_{k=0}^{m-1} \lambda_k \ge \sum_{k=0}^{m-1} \sigma_k^2 \qquad \text{Ec 4. 25}$$

para $1 \le m \le M$.◊

Teorema 4. Minimización de la función distorsión-tasa de velocidad. La transformada KL tiene la mínima tasa de distorsión-velocidad entre todas las transformaciones unitarias, es decir

$$R(\Phi^{*T}) \le R(A) \qquad \text{Ec 4. 26}$$

Demostración. Aquí se asume que los vectores x,y,$\tilde{x}$,$\tilde{y}$ son gaussianos, entonces

$$D = \frac{1}{N} E[(x - \tilde{x})^{*T}(x - \tilde{x})] \qquad \text{Ec 4. 27}$$

como T es ortogonal y, $x = T^{*T} y$, $\tilde{x} = T^{*T} \tilde{y}$, se tiene que

$$D = \frac{1}{N} E[(y - \tilde{y})^{*T} T T^{*T} (y - \tilde{y})] = \frac{1}{N} E[(y - \tilde{y})^{*T} (y - \tilde{y})] = \frac{1}{N} E[\delta y^{*T} \delta y] \qquad \text{Ec 4. 28}$$

donde $\delta y = y - \tilde{y}$ representa el error al reproducir y. Como D es invariante, se tiene que

$$R = \frac{1}{N} \sum_{k=0}^{N-1} \text{máx } [0, \frac{1}{2} \log_2 \frac{\sigma_k^2}{\theta}] \qquad \text{Ec 4. 29}$$

$$D = \frac{1}{N} \sum_{k=0}^{N-1} \text{máx } [\theta \sigma_k^2] \qquad \text{Ec 4. 30}$$

como $\sigma_k^2 = E[|y(k)|^2] = [TRT^{*T}]_{k,k}$, $R = R(T)$ es mínima cuando $T = \Phi^{*T}$, entonces

$$R(\Phi^{*T}) \leq R(A) \qquad \text{Ec 4. 31}$$

lo que concluye la demostración.◊

4.2. Obtención de la KLT

Después de las etapas anteriores, se han obtenido vectores de bandas críticas por trama de dimensión 1x18, con valores no negativos. También segmentos como conjuntos de N tramas de vectores de bandas críticas. Para estos segmentos se obtendrá la representación KLT.

Las matrices de covariancias tienen las características de ser simétricas, reales, y no negativas. También son positivas semidefinidas[68].

Teorema 5. La matriz de covariancias es semipositiva definida. Sea C una matriz de covariancias, entonces para cualquier vector x, se tiene que

$$x^T Cx \geq 0. \qquad \text{Ec 4. 32}$$

Demostración. Sea C la matriz de covariancias del vector aleatorio z de media μ_z. Como es claro observar,

$$E\left\{\left[x^T(z-\mu_z)\right]^2\right\} \geq 0 \qquad \text{Ec 4. 33}$$

de donde,

$$E\left\{\left[x^T(z-\mu_z)\right]^2\right\} = z^T E\left[(z-\mu_z)(z-\mu_z)^T\right]z = z^T Cz \geq 0 \qquad \text{Ec 4. 34}$$

lo que completa la demostración.◊

Más aún, con probabilidad uno la matriz de covariancias en el caso de vectores de bandas críticas en subpalabras es de rango completo, y en consecuencia, es positiva definida. Entonces usando el siguiente teorema se asegura la positividad de los valores propios[69] de matrices de covariancias para subpalabras.

Teorema 6. Positividad de los valores propios. Sea A una matriz simétrica de orden n, si A es positiva definida entonces los valores propios de A son positivos.

Demostración. Sea N el orden de $A, M \leq N, \lambda$ un valor propio de la submatriz principal $A(M,M)$, y $x(M)$ un vector propio de $A(M,M)$ que corresponde a λ.

Sea x un vector columna que se obtiene al añadir ceros al vector $x(M)$ en los índices fuera de M. Entonces $(Ax,x) = (A(M,M)x(M),x(M))$. Ya que el miembro izquierdo es positivo la parte derecha puede expresarse como $(\lambda x(M),x(M)) = \lambda(x(M),x(M))$. Como $(x(M),x(M))$ es positivo, entonces λ es positivo.

Ahora $\det A(M,M)$ es igual al producto de todos los valores propios de $A(M,M)$, como estos son positivos, $\det A(M,M) > 0$, de donde todos los menores principales de A son positivos. Del polinomio propio

$$\sum^{N} \lambda^k c_{N-k} = 0 \qquad \text{Ec 4. 35}$$

donde c_k es la suma de los k menores principales de A y $c_0 = 1$, esta suma es siempre positiva, de donde se observa que λ es positiva.◊

Dada una matriz de covariancias C_{xx} de orden n, que incluye una o más tramas, la KLT se define por tres elementos: Un conjunto de n valores propios, un conjunto de n vectores propios de dimensión n, y un vector de coeficientes de dimensión n por cada vector de banda crítica.

El siguiente teorema asegura la existencia de los dos primeros elementos, y por ende del último, en el caso de matrices de covariancias.

Teorema 7. Diagonalización de una matriz simétrica. Si A es una matriz simétrica de orden n, entonces:

1. La matriz A es diagonalizable.
2. Los valores propios de A pertenecen a $\Re$.
3. Existe una matriz ortogonal P tal que $D = P^{-1}AP$, donde D es una matriz diagonal formada por valores propios de A

4. P es una matriz formada por vectores ortonormales asociados a los valores propios correspondientes, por lo que $D = P'AP$.

5. $T(A) = P'AP$ preserva el espectro, esto es, $\det(A) = \det(D)$ y $tr(A) = tr(D)$.

La demostración de este teorema es muy conocida. Si además A es no-negativa, entonces, P es también una matriz positiva permutacional generalizada. esto implica que D es también una matriz no-negativa.

Dados los vectores de bandas críticas por trama, se obtiene su matriz de correlaciones muestral en una subpalabra, usando una estimación aritmética, y se resta a la matriz de medias en ese segmento, esto es

$$C_{XX} = R_{XX} - \overline{X} \qquad \text{Ec 4. 36}$$

De la matriz de covariancias C_{XX} se obtienen los elementos KLT.

Los coeficientes de la KLT son vectores de dimensión J, donde J es el número de bandas críticas con M vectores, donde M es el número de tramas en la subpalabra. Los coeficientes proporcionan información temporal de las variaciones de los vectores en el segmento. Si consideramos que los segmentos son cuasiestacionarios, entonces los coeficientes reportan variaciones lentas en sus valores, que tienen menor interés para el reconocimiento.

Los vectores propios de un segmento nos dan información del segmento en su conjunto, de hecho a lo largo de estos vectores se pueden giran los ejes coordenados, y representan variaciones críticas del segmento. Estos vectores caracterizan al segmento, aún más si se conservan sólo alrededor de los siete primeros vectores, son suficientes para discriminar fonemas parecidos.

Al conservar en una subpalabra los valores propios y solo siete vectores propios, se tiene un conjunto de 8 vectores de dimensión 18 que representan a la subpalabra, ésta es una compresión importante para el reconocimiento, ya que en una subpalabra existen de 50 a 100 tramas, de manera que en un sistema clásico LPC o cepstral, se tienen de 50 a 100 vectores de dimensión alrededor de 8.

Los vectores propios se conservan en la etapa de entrenamiento y se añaden los valores propios para las palabras de prueba, de manera que en la etapa de clasificación, la energía utilizada es la de los vectores de prueba.

Como puede esperarse es necesario utilizar una medida de Mahalanobis para comparar subpalabras a través de vectores y valores propios, esto es,

$$d(x,y)=(x-y)^T W(x-y) \qquad \text{Ec 4. 37}$$

en este caso, la matriz de pesos W es diagonal ya que estaría formada por valores propios, por lo que

$$d(x,y)=\sum_{i=1}^{k} w_{ii}(x_i-y_i)^2 \qquad \text{Ec 4. 38}$$

La medida utilizada es semejante a ésta[45], se establece como

$$d=\frac{\lambda^T(I-A^TA)\lambda}{n} \qquad \text{Ec 4. 39}$$

donde λ es el vector de valores propios ordenados de mayor a menor, n es el número de vectores y valores tomados, y A es la matriz de productos internos entre los vectores propios.

La utilización de las medidas planteadas en las ecuaciones 4.38 y 4.39 proporcionan resultados semejantes durante el reconocimiento, siendo indistinta su aplicación en los sistemas planteados.

4.3 La transformada KL en señales de voz

Las aplicaciones de la KLT a palabras muestran una transformada robusta para segmentos acústicos o de otro tipo, que incluyan inclusive a más de 100 tramas. En esta sección se presentan algunos ejemplos de los parámetros KLT para palabras tanto en español como en inglés, para distinto número de segmentaciones.

En la figura 4.1 se muestra dos repeticiones de la palabra *one* tomando toda la palabra como todo un segmento, que sería el caso más complicado para parametrizar.

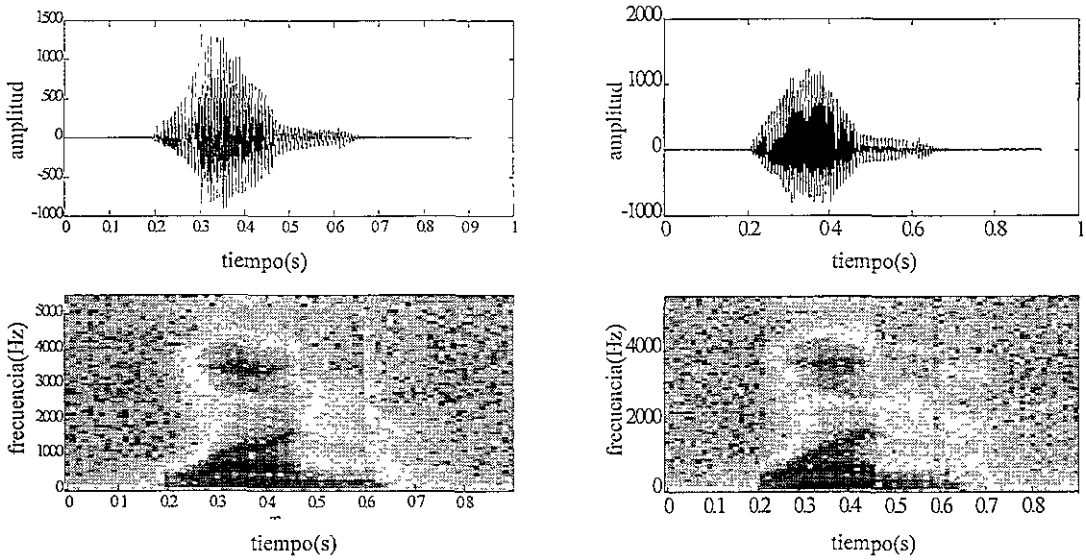


Fig. 4.1. Dos repeticiones de la palabra en inglés *one*, donde se muestran las señales en el tiempo y sus espectogramas.

Para las repeticiones de la figura 4.1 se muestran en la figura 4.2 los primeros vectores propios, incisos (a) y (b), los segundos vectores propios, (c) y (d), así como los vectores propios en los incisos (e) y (f).

En la figura 4.2 puede observarse la información espectral en los primeros vectores. En estos predomina una amplitud fuerte en bajas frecuencias y en alguna franja destacable en medias frecuencias. En los valores propios, la caída de estos a partir del quinto reiteran el predominio del fonema vocálico en las repeticiones.

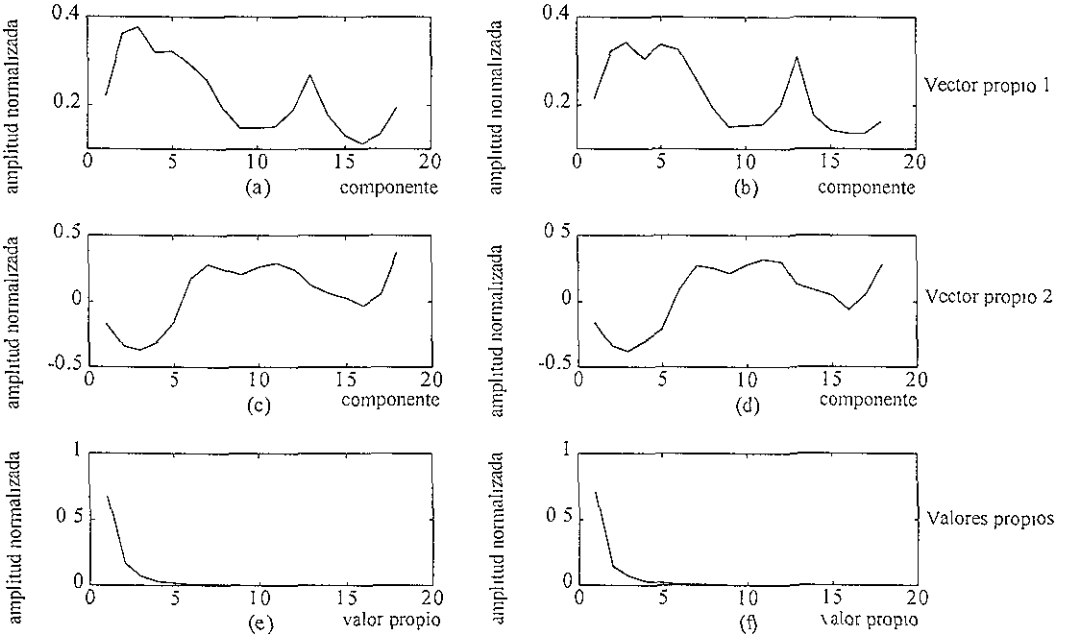


Fig. 4.2. Los dos primeros vectores propios de dos repeticiones de la palabra en inglés *one*, primeros vectores (a) y (b), segundos (c) y (d), y los valores propios, incisos (e) y (f).

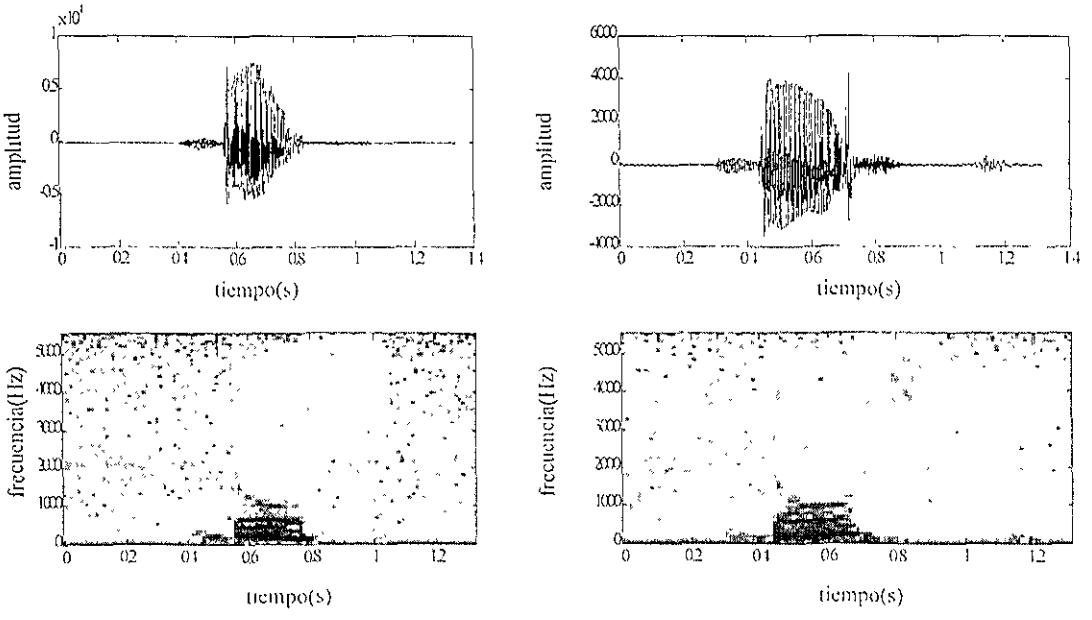


Fig. 4 3 Dos repeticiones de la palabra en español *dos*, donde se muestran las señales en el tiempo y sus espectogramas

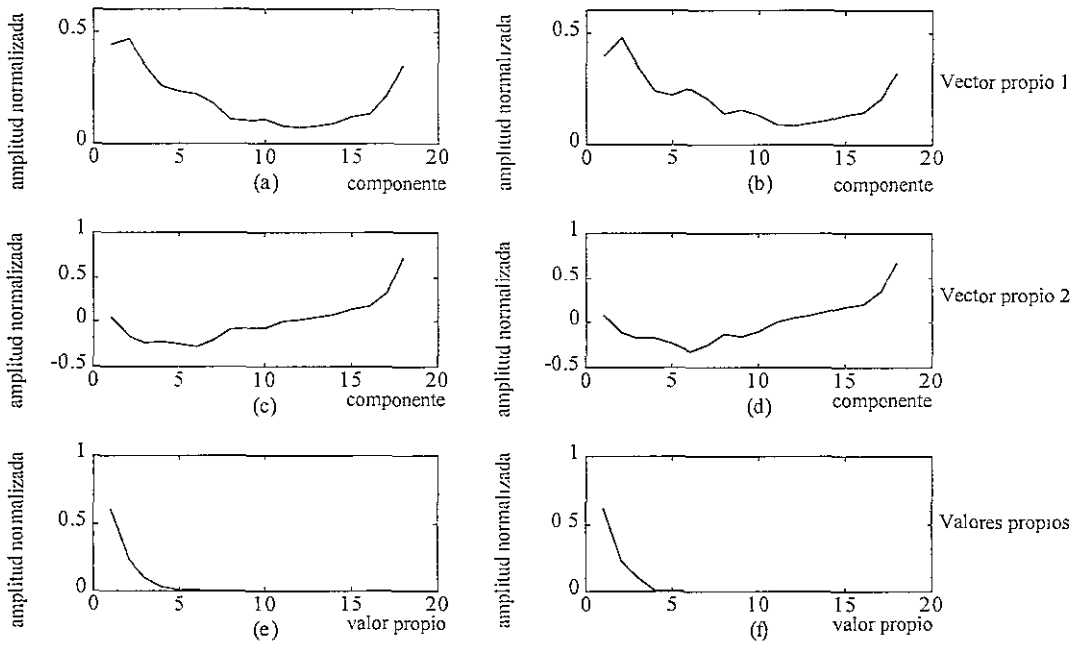


Fig. 4.4. Los dos primeros vectores propios de dos repeticiones de la palabra en español *dos*, primeros vectores (a) y (b), segundos (c) y (d), y los valores propios, incisos (e) y (f).

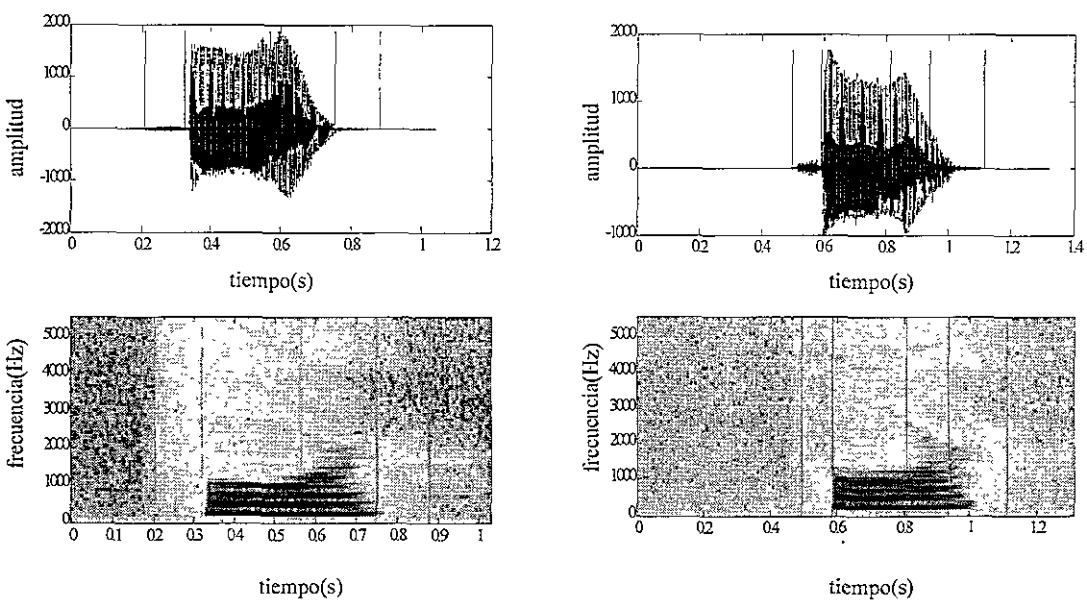


Fig. 4.5. Dos repeticiones de la palabra en inglés *four*, donde se muestran las señales en el tiempo y sus espectogramas, con 4 segmentos acústicos.

De manera complementaria se presentan en las figuras 4.3 y 4.4 el mismo tipo de gráficas para la palabra *dos*, donde se observa una mayor similitud en los parámetros KLT que para la palabra en inglés *one,* resultado de una mayor similitud en sus espectros.

En la figura 4.5 se muestran dos repeticiones de la palabra en inglés *four*, tanto la señal en el tiempo y su espectrograma, con cuatro segmentos acústicos. El primer segmento incluye el fonema /f/ los dos siguientes parten el diptongo */ou/* y el último segmento corresponde a la vibrante /r/.

En la figura 4.6 se muestran los vectores propios 1 y 2, así como los valores propios correspondientes al primer segmento. En las primeras gráficas se observa la amplitud acentuada en altas frecuencias, en la segunda repetición se amplía a frecuencias medias, esto corresponde a su comportamiento espectral, figura 4.5.

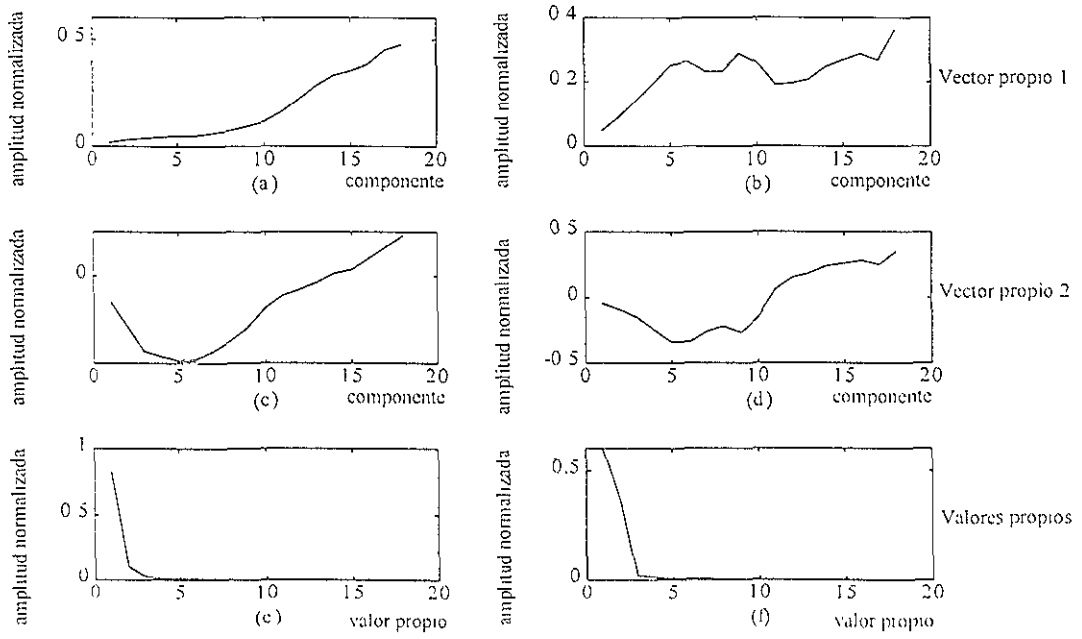


Fig. 4.6. Los dos primeros vectores propios de dos repeticiones de la palabra en inglés *four*, primeros vectores (a) y (b), segundos (c) y (d), y los valores propios, incisos (e) y (f), primer segmento

En la figura 4 7 se muestran los mismos parámetros KLT para el cuarto segmento. Se han escogido el primer y cuarto segmento, ya que son los de mayor interés para palabra *four* En

el cuarto segmento, los vectores propios se observan más semejantes, esto procede de un comportamiento espectral también más parecido. En ambos segmentos los valores propios tienden a cero rápidamente, a pesar de que sus espectros contienen amplitudes en medias y altas frecuencias superiores a las de bajas frecuencias. Este hecho, consistente en que las palabras utilizadas nos llevan a tomar una cota heurística de 7 valores y vectores propios, suficientes para describir adecuadamente un segmento. En términos de energía, se dio que estos 7 valores incluían más de 97% en la mayoría de los casos y más de 95% en pocas repeticiones de fricativas.

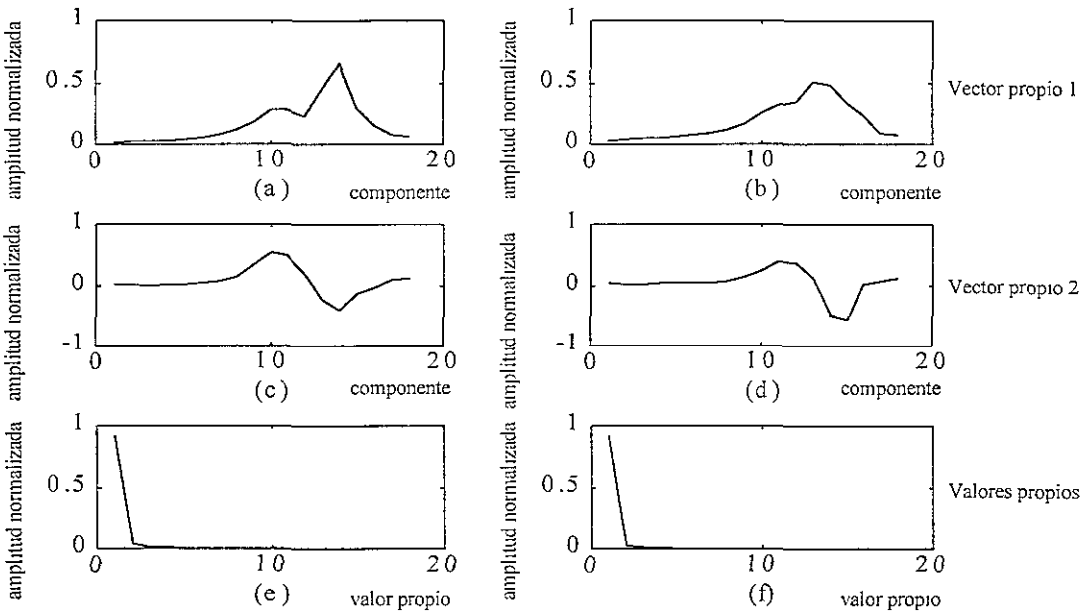


Fig. 4.7. Los dos primeros vectores propios de dos repeticiones de la palabra en inglés *four*, primeros vectores (a) y (b), segundos (c) y (d), y los valores propios, incisos (e) y (f), cuarto segmento.

En la figura 4.8 se muestra la señal en el tiempo y el espectro de dos repeticiones de la palabra *cinco*, dividida en 4 segmentos acústicos. El primer segmento corresponde al fonema /s/, que en español pierde sus características fricativas, el segundo segmento ocupa los fonemas /i/ y /n/ así como parte del silencio del fonema oclusivo /k/. Como se ha fijado la segmentación en 4 segmentos, el método selecciona los cambios más fuertes, así que a pesar de que en el segundo segmento incluyen variaciones espectrales significativas no se produce un corte. El último segmento corresponde en realidad a ruido producido por la terminación del fonema /o/.

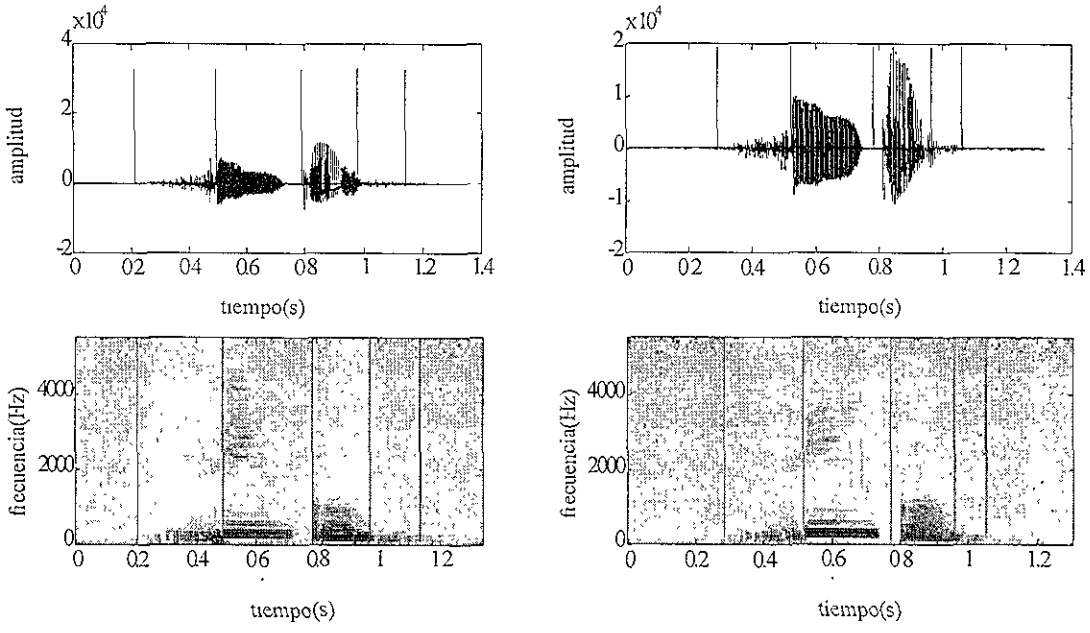


Fig. 4.8. Dos repeticiones de la palabra en español *cinco* donde se muestran las señales en el tiempo y sus espectogramas, cuatro segmentos.

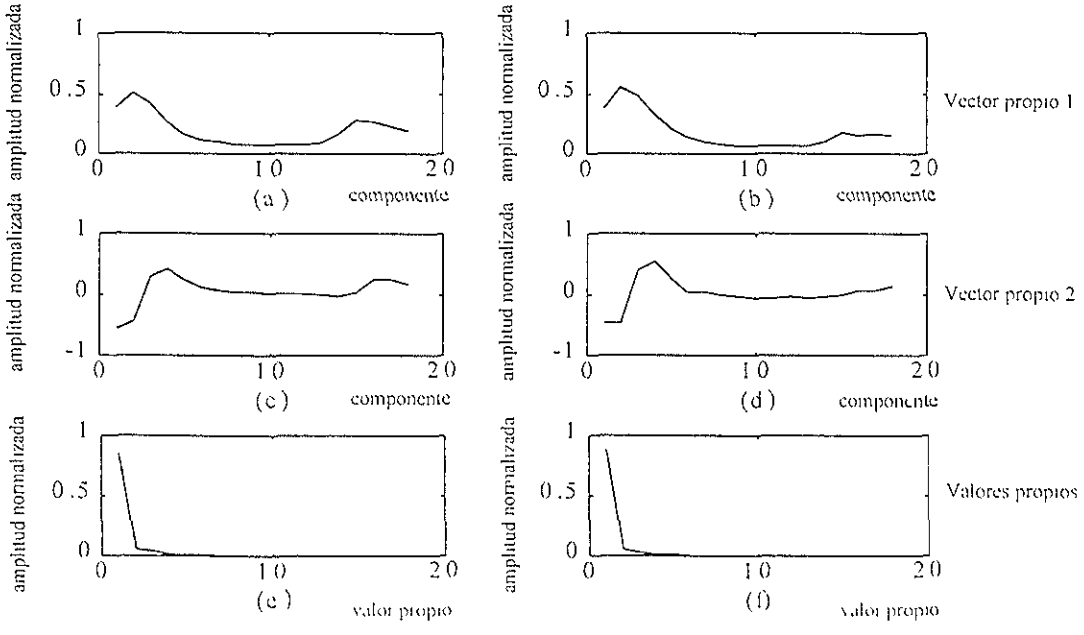


Fig. 4 9. Los primeros dos vectores propios de dos repeticiones de la palabra en español *cinco* en el primer segmento, (a) y (b) primeros vectores , (c) y (d) segundos , e incisos (e) y (f) los valores propios

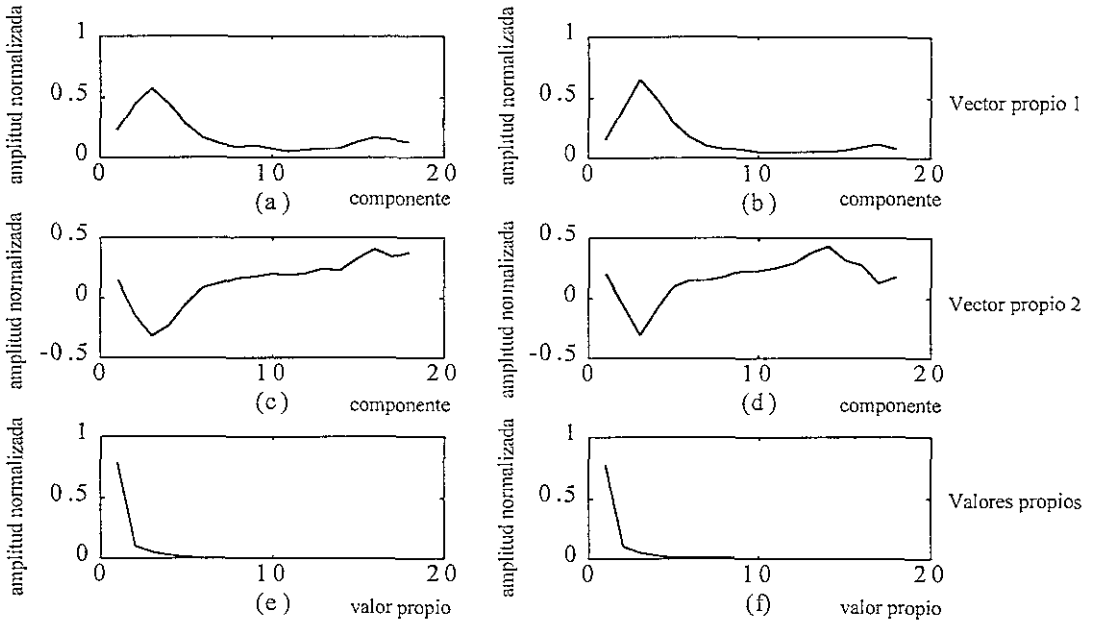


Fig. 4.10. Los primeros dos vectores propios de dos repeticiones de la palabra en español *cinco* en el segundo segmento, (a) y (b) primeros vectores , (c) y (d) segundos , e incisos (e) y (f) los valores propios.

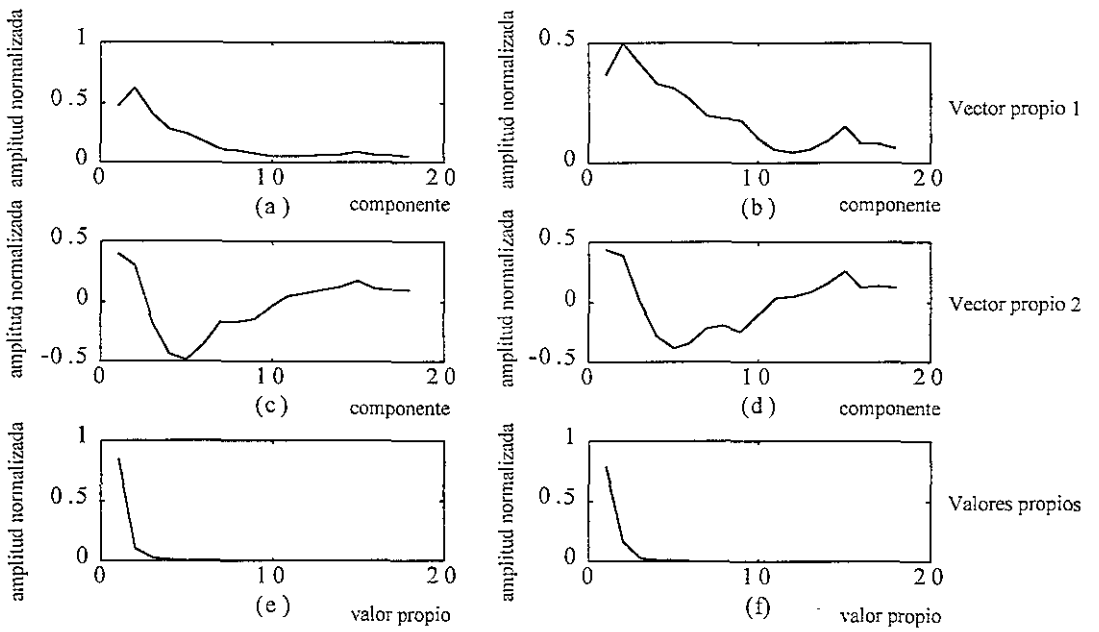


Fig. 4.11. Los primeros dos vectores propios de dos repeticiones de la palabra en español *cinco* en el tercer segmento, primeros vectores (a) y (b), segundos (c) y (d), y los valores propios, incisos (e) y (f), .

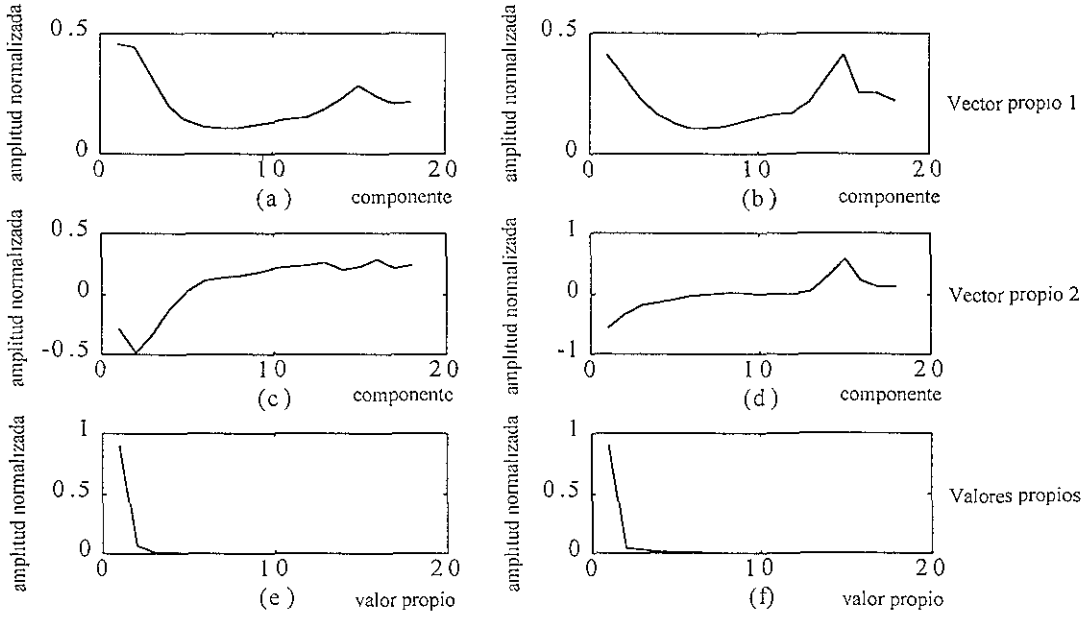


Fig. 4.12. Los primeros dos vectores propios de dos repeticiones de la palabra en español *cinco* en el cuarto segmento, (a) y (b) primeros vectores , (c) y (d) segundos , e incisos (e) y (f) los valores propios.

En las figuras 4.9 a 4.12 se han graficado los primeros dos vectores propios y los valores propios de dos repeticiones de la palabra *cinco* por subpalabra, se hizo de esta forma para observar el comportamiento de estos parámetros en todas las subpalabras y verificar la robustez de estos en cualquier tipo de fonema.

En las figuras 4.9 a 4.12 se puede observar la información espectral que contiene el primer vector propio. Los vectores propios subsecuentes proporcionan también información crítica del comportamiento del segmento, aunque no de manera tangible como el caso del primer vector propio[45][70]

Capítulo 5

CLASIFICACIÓN DE PALABRAS AISLADAS

El objetivo de los sistemas automáticos de reconocimiento de voz es contar con un transductor automático de voz continua a texto, solo es cuestión de tiempo para que se pueda contar con éste. La voz es el medio incuestionable de la comunicación humana, al ofrecer manos y ojos libres que otros medios de comunicación no tienen. No sólo se obtienen las ventajas inherentes a la comunicación humana sino también a la necesidad de comunicarse eficientemente con las máquinas, la comunicación por voz se caracteriza por su sencillez, conveniencia y rapidez.

La historia del procesamiento de voz data de fines de los 40. Muchas técnicas han sido elaboradas desde entonces, hasta hoy se han comercializado sistemas de reconocimiento de palabras continuas con vocabularios pequeños, mucho gracias al desarrollo del hardware en procesadores digitales de señales. Actualmente se estudian los sistemas de reconocimiento de palabras continuas en laboratorios, sin embargo, parece que tardará aún algún tiempo antes de que las máquinas puedan reconocer palabras de una conversación común y corriente. Las complicaciones existen no sólo del hardware, sino también para utilizar la información lingüística y contextual.

Los sistemas de reconocimiento basados en segmentación matemática utilizan distorsión dinámica en tiempo (DTW) y modelos ocultos de Markov. Ambos algoritmos emplean principios dinámicos de programación para realizar alineamientos no lineales en el tiempo. Mientras los algoritmos DTW utilizan formas deterministas para la programación dinámica, los sistemas HMM emplean una forma aleatoria de programación dinámica conocida como algoritmo de Viterbi.

Todos los sistemas de reconocimiento incluyen dos fases: entrenamiento, y reconocimiento o clasificación. La fase de entrenamiento establece una memoria de referencia o diccionario de patrones de voz a los cuales se les asigna etiquetas lingüísticas. En la fase de reconocimiento, se asigna una etiqueta a una repetición de prueba, obviamente desconocida.

La labor del reconocimiento de voz es fundamentalmente una tarea de clasificación de patrones. El objetivo es identificar una palabra o conjunto de éstas de un vocabulario de palabras almacenadas en memoria y parametrizadas convenientemente. La clasificación es una comparación entre una señal desconocida y un conjunto de señales de referencia. Las referencias se derivan de una fase de entrenamiento preliminar, previa al reconocimiento. En el reconocimiento, la señal por clasificar es comparada con todas las referencias almacenadas en memoria o un subconjunto de éstas, esto último tiene por objetivo mejorar los procedimientos de búsqueda. La comparación involucra una medida de distancia que analice la similaridad entre las señales de prueba y referencia. La referencia más cercana a la de prueba es la seleccionada como la señal reconocida.

Los sistemas de reconocimiento se clasifican en cuatro tipos básicos: sistemas determinísticos basados en cuantización vectorial (VQ) y otros en ajuste dinámico en el tiempo (DTW), sistemas estadísticos como los modelos ocultos de Markov (HMM), y redes neuronales (NN). Para el reconocimiento de palabras continuas, todos los sistemas anteriores incluyen una fase de información lingüística auxiliar indispensable.

5.1. Clasificación usando DTW

5.1.1. Alineamiento dinámico en el tiempo

Si se observa una señal de voz en el tiempo o su propio espectograma se podrán observar las variaciones no lineales, tanto en amplitud como en tiempo para repeticiones de una palabra o un conjunto de éstas.

Una característica de los parámetros como LPC, cepstral u otros de tramas de señales de voz es que eliminan o disminuyen variaciones en amplitud de las diferentes repeticiones; sin embargo, las variaciones en tiempo permanecen ya que el análisis es por tramas cortas. Las variaciones temporales para repeticiones de una misma señal de voz han sido un problema crucial en el reconocimiento automático de palabras.

En 1975, Itakura[7] propuso un método de normalización no lineal en el tiempo, basado en una función de ajuste y una técnica de programación dinámica. Para el diseño de la función de ajuste se basó a su vez en la propuesta de Velichiko y el método de programación dinámica fue propuesto originalmente por Sakoe y Chiba[18]. En este algoritmo, las fluctuaciones de tiempo son modeladas con una función no lineal con propiedades específicas.

Considérese el problema de eliminar las variaciones de tiempo entre dos vectores parametrizados $A = a_1, a_2, \ldots, a_M$ y $B = b_1, b_2, \ldots, b_N$ de M y N tramas cada uno, que corresponden a repeticiones de prueba y referencia respectivamente, como se ilustra en la figura 5.1.

En la figura 5.1 se observa que DTW establece una correspondencia entre cada vector de A al vector correspondiente de B más cercano, la función de correspondencia resultante C es llamada función de ajuste. Cada valor de la función $c(k)$ está formada por un par de índices $c(k) = [i(k), j(k)]$.

El error entre cada punto de la función de ajuste está dada por la distancia

$$D(A, B) = \min_{c} \sum_{k=1}^{K} d(c(k)) \qquad \text{Ec 5. 1}$$

En la figura 5.1 se observa por ejemplo que los puntos inicial y final pueden no coincidir, o que existen límites al ajuste para evitar excesivas diferencias en tiempo. En general, se han establecido condiciones a la función de ajuste de manera experimental, entre las más conocidas e importantes se tiene las siguientes:

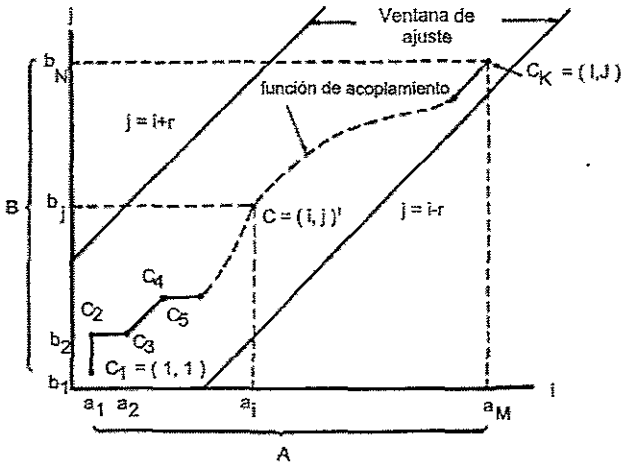


Figura 5.1. Función de ajuste DTW

- **De variación monotónica creciente.** Se refieren a la variación monotónica creciente no estricta entre puntos consecutivos para ambos vectores por clasificar, esto es,

$$i(k-1) \le i(k) \qquad j(k-1) \le j(k) \qquad \text{Ec 5. 2}$$

Éstas son las condiciones más universalmente aceptadas, y no existen sistemas que establezcan lo contrario. Para palabras no se tiene porque establecer otro tipo de variación.

- **De continuidad por intervalos.** Se puede tener una continuidad en la consideración de puntos en el ajuste, es decir

$$i(k) - i(k-1) \le 1 \qquad j(k) - j(k-1) \le 1 \qquad \text{Ec 5. 3}$$

o bien, una continuidad por intervalos al establecer un límite p en los puntos que no se tomarían en consideración, esto es,

$$i(k) - i(k-1) \leq p \qquad j(k) - j(k-1) \leq p \qquad \text{Ec 5. 4}$$

este último tipo de continuidad permite que se eliminen tramas de ruido entre fonemas que dificultan el reconocimiento, o bien fonemas sordos fricativos muy débiles que no aparecen en ciertas repeticiones. Sin embargo, el límite p debe diseñarse cuidadosamente para no conducir a clasificaciones erróneas en palabras similares.

- **De puntos inicial y final**. La forma más simple de considerar los puntos inicial y final, es corresponderlos, esto es,

$$i(1) = 1, \quad j(1) = 1, \quad i(K) = M, \quad j(K) = N \qquad \text{Ec 5. 5}$$

sin embargo, la forma más general es dejar variaciones, m_1, m_2, n_1, n_2, en los puntos inicial y final,

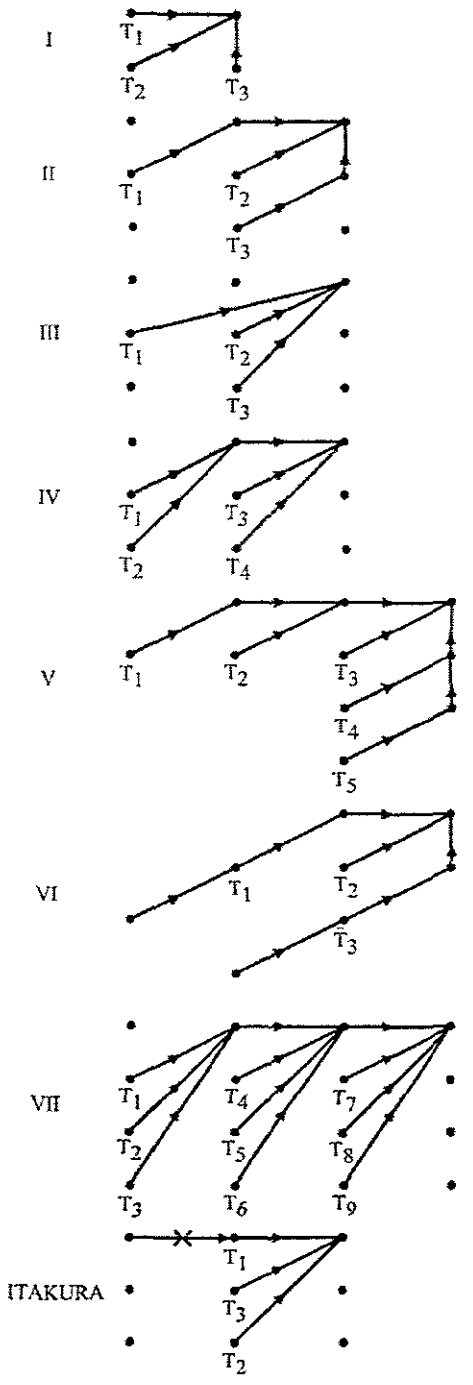
$$i(1) \leq m_1, \quad j(1) \leq m_2, \quad i(K) \leq M - n_1, \quad j(K) \leq N - n_2 \qquad \text{Ec 5. 6}$$

esta forma conduce a eliminar ruido al inicio o final de la palabra.

- **Trayectorias locales**. Las trayectorias conducen a limitar los pasos consecutivos ya sean horizontales o verticales, las compresiones requieren que en el contorno se lleven a cabo no más de m pasos horizontales o verticales sin hacer primero los n pasos diagonales Estas trayectorias validan también los movimientos de puntos consecutivos, como puede observarse en la figura 5.2 en la que se muestran 8 trayectorias comúnmente usadas por Sakoe y Chiba, tipos I, IV, V y VII[18] ; por Itakura, tipo III[7], donde × indica una ruta prohibida; y por Myers, tipos II, III, IV, VI, VIII[20]. No tienen un patrón común o metodología, sino que fueron establecidas experimentalmente. En éstas se observa la no continuidad en las trayectorias III, VI y VIII.

  Tanto Sakoe, Chiba como Myers *et al.* evaluaron las diferentes posibilidades utilizándolas en sistemas de reconocimiento de palabras y comparando las precisiones del reconocimiento. Sakoe y Chiba obtuvieron el mejor desempeño para el tipo I et al,

Myers et al, comparó los tipos II, III, IV, VI y VIII con desempeños semejantes, pero el tipo III se comportó significativamente peor.

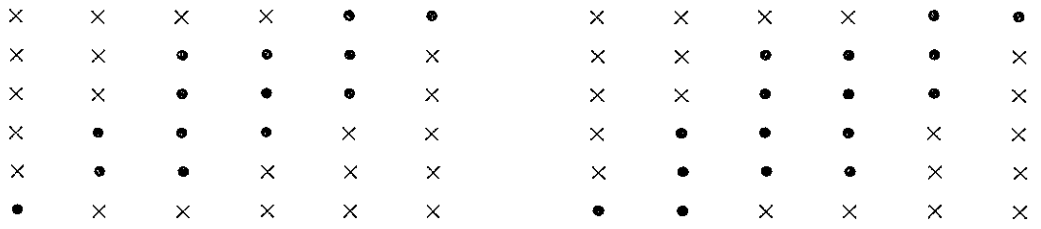


Fgura 5.2. Trayectorias locales consideradas por varios autores[32].

- **Límites globales.** Las trayectorias locales conducen a trayectorias globales que reducen la región de puntos permitidos para trayectorias de ajuste, estas trayectorias

alrededor de la ruta lineal forman generalmente polígonos y no paralelogramos como alguna literatura propone[32][71].

Así para trayectorias tipo I, se tiene que todos los puntos de los vectores por acoplar son permitidos. En la figura 5.3 se muestran los límites globales para trayectorias locales tipos IV y de Itakura, para acoplamiento de conjuntos de 6 vectores de dimensión unitaria en ambos ejes.



(a) Tipo IV (b) Tipo Itakura

Figura 5.3. En los límites globales de 6x6 puntos, se muestran con círculos los puntos permisibles y con cruces los nos permitidos, uniendo puntos (1,1) y (6,6) de trayectorias tipo (a) IV y (b) Itakura .

Se pueden establecer trayectorias globales independientes de las trayectorias locales. En particular, ha sido utilizada la propuesta por Sakoe y Chiba[18], $|x-y|$, que conduce a puntos de inicio y fin variables.

- **Pesos a la pendiente**. Una función de peso se aplica a cada distancia, con objeto de privilegiar ciertas trayectorias locales. Se usan en adición a éstas o además de las trayectorias Nuevamente Sakoe y Chiba[18] propusieron las siguientes

$$\begin{aligned} m(k) &= \min[x(k)-x(k-1), y(k)-y(k-1)] \\ m(k) &= \max[x(k)-x(k-1), y(k)-y(k-1)] \\ m(k) &= x(k)-x(k-1) \\ m(k) &= [x(k)-x(k-1)]+[y(k)-y(k-1)] \end{aligned} \qquad \text{Ec 5. 7}$$

En consideraciones prácticas de DTW para reconocimiento de voz, las rutas tienden a ramificarse en exceso y el número de posibilidades tiende a crecer de esta forma. Teóricamente, DTW requiere el cálculo de $I \times J$ tramas de distancia para entre dos repeticiones, donde I y J son las longitudes de las señales de prueba y de referencia

respectivamente. En la práctica, las consideraciones globales y locales restringen la búsqueda de tal modo que sólo 25% a 35% de las combinaciones se llevan a cabo. Sin embargo, existe un incremento substancial de cálculos en relación con una comparación lineal trama por trama. Por ejemplo, para una señal de 250 tramas, una función de ajuste lineal requiere de 250 cálculos de distancia, mientras que típicamente DTW requiere de 2,500 a 3,500 cálculos. Los cálculos basados en DTW son significativamente grandes para secuencias de palabras.

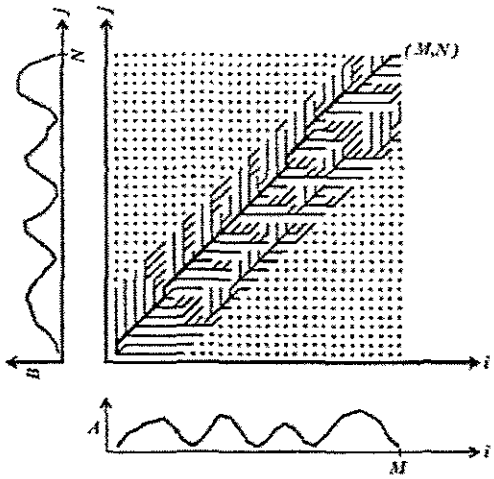


Figura 5.4. Rutas de ajuste en el tiempo, la óptima se señala por la línea sombreada.

También es pertinente añadir la imposición de un máximo permisible $D(C)$, en donde las rutas que por no ajustarse exceden el límite se ignoran muy pronto en la búsqueda. Los candidatos resultantes típicamente forman un paquete de rutas, como se muestra en la figura 5.4.

Al final del proceso, la ruta correcta se encuentra comenzando en (M,N) y retrocediendo hasta el punto inicial; en la figura 5.4, la ruta óptima está indicada por la ruta sombreada.

El cálculo de la función de ajuste es equivalente a encontrar la ruta de costo mínimo a través de una rejilla de puntos sujetos a un conjunto de rutas y límites de fin de puntos. Este problema de minimización se resuelve eficientemente en DTW utilizando algoritmos de programación dinámica. El concepto de programación dinámica ha sido empleada exitosamente para resolver problemas de optimación que puedan ser decompuestos en una secuencia de optimaciones parciales, en los cuales solo existe un número límite de posibles opciones de rutas en cada paso de optimación.

Basado en este principio, se establece la siguiente búsqueda recursiva. Si $D(c(k))$ se define como la distancia acumulativa óptima al punto $c(k)$ y $D(c_{k-1})$ la distancia óptima al punto $k-1$, entonces el algoritmo está dado por:

$$D(c(k)) = \min_{legal[c(k-1)]} \{d[c(k)] + [D(c_{k-1})]\} \qquad \text{Ec 5. 8}$$

donde $legal[c(k-1)]$ define el conjunto de predecesores permisibles de $c(k-1)$ a $c(k)$.

En general, el algoritmo está dado por:

1. Inicio:

$$\begin{aligned} D_1(i,n) &= d(c(i,n)) \\ \varepsilon_1(n) &= i \\ para \quad n &= 1,2,\ldots,N \end{aligned} \qquad \text{Ec 5. 9}$$

2. Proceso recursivo

$$\begin{aligned} D_{m+1}(i,n) &= \min_{1\le l\le N}[D_m(i,l) + d(l,n)] \\ \varepsilon_{m+1}(n) &= \arg\min_{1\le l\le N}[D_m(i,l) + d(l,n)] \\ para \quad n &= 1,2\ldots,N \quad y \quad m = 1,2,\ldots,M-2 \end{aligned} \qquad \text{Ec. 5.10}$$

3. Terminación

$$\begin{aligned} D_M(i,j) &= \min_{1\le l\le N}[D_{M-1}(i,l) + d(l,j)] \\ \varepsilon_M(j) &= \arg\min_{1\le l\le N}[D_{M-1}(i,l) + d(l,j)] \end{aligned} \qquad \text{Ec.5 11}$$

4. Búsqueda de trayectoria

$$\begin{aligned} &\text{trayectoria óptima} = (i, i_1, i_2, \ldots, i_{M-1}, j), \\ &\text{donde} \quad i_m = \varepsilon_{m+1}(i_{m+1}) \quad m = M-1, M-2, \ldots 1 \quad con \quad i_M = j. \end{aligned}$$

Los mayores problemas al utilizar DTW son (a) una carga pesada de cálculos y, (b) incapacidad para explotar información temporal de la señal Aún más la estructura básica de DTW no permite conocer la contribución de diferentes partes de una frase al reconocimiento.

5.1.2. Método de clasificación

En los métodos hasta ahora utilizados, los algoritmos DTW se han aplicado a tramas cortas dentro de una palabra o un conjunto de éstas. En este trabajo, la clasificación se diseña y utiliza el DTW en el nivel de subpalabras acústicas. Se han experimentado distintos tipos de trayectorias locales, la que mejor se ajusta es una variación de Itakura sin restricciones horizontales, no hay pesos adicionales y tenemos trayectorias globales adicionales para introducir puntos inicial y final variables, con un máximo de una subplabra de corrimiento al inicio y dos al final en ambos ejes. Esta última consideración es crucial en el reconocimiento de palabras que inician o terminan en fricativas débiles.

Para la clasificación DTW, es posible utilizar la misma base de datos de palabras en entrenamiento que en clasificación, obviamente eliminando la repetición que se está probando de la base de entrenamiento, ver figura 5.5. En el modo de entrenamiento, se almacenan por palabra, 7 vectores propios de cada supalabra de dimensión 1x18.

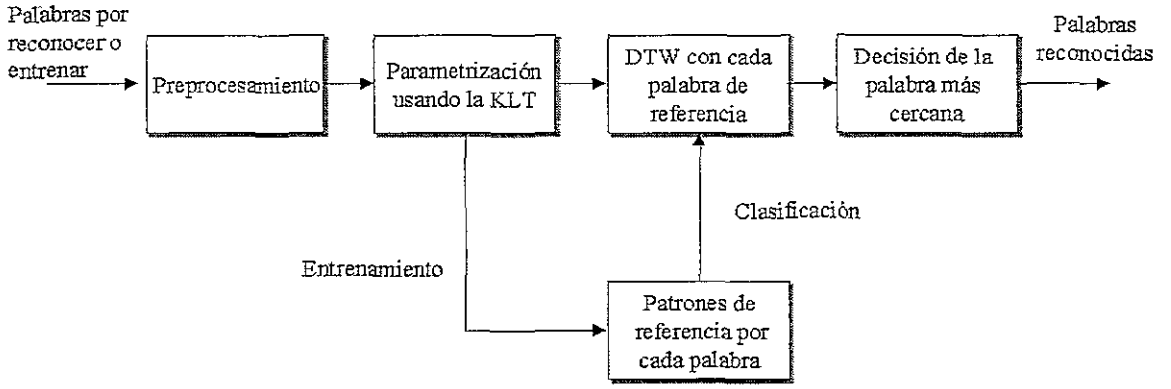


Figura 5.5. Diagrama de bloques del sistema de reconocimiento de palabras aisladas usando DTW.

Se ha evaluado el rendimiento usando primeramente la base de datos de TI, se usaron los dígitos para 5 hombres y 5 mujeres, con 10 repeticiones de cada quien para entrenamiento, de forma tal que se tienen 1 000 repeticiones en esta fase. Para la clasificación se utilizaron 1 600 repeticiones distintas alcanzando una precisión de 99.25%, como se muestra en la matriz de confusión , ver tabla 5.1.

En una primera instancia se había obtenido una precisión de 95.6% usando un método semejante al reportado por el tutor[46], que utilizaba trayectorias locales del tipo I, uniendo principios y fines de palabra, y trayectorias globales de pendiente ½. Sin embargo, se rehizo el algoritmo MLR teniendo un mayor cuidado en el inicio y fin, donde se combinó el método MLR con el método Rabiner y Sambur, así como permitiendo corrimientos de inicio y fin con lo que se logró un rendimiento de 98.4%[72], ver tabla 5.2. Finalmente, al variar el tipo de trayectoria a la mencionada y ajustar corrimientos de inicio y fin de palabra, se logró elevar el rendimiento al presentado en la tabla 5.1.

	0	1	2	3	4	5	6	7	8	9
0	159	0	0	0	0	0	0	0	0	0
1	1	157	0	0	0	0	0	0	0	2
2	0	0	158	2	0	0	0	0	0	0
3	0	0	2	158	0	0	0	0	0	0
4	1	0	0	0	159	0	0	0	0	0
5	0	0	0	0	0	160	0	0	0	0
6	0	0	0	1	0	0	159	0	0	0
7	0	0	0	0	0	0	0	160	0	0
8	0	0	0	0	0	1	0	0	159	0
9	0	1	0	0	0	0	0	0	0	159

Tabla 5.1. Matriz de confusión para palabras aisladas en inglés de la base de TI, método DTW, con parámetros KLT, trayectorias modificadas, precisión 99.25%.

	0	1	2	3	4	5	6	7	8	9
0	157	1	0	1	0	0	0	0	0	1
1	1	155	0	0	1	0	0	0	0	3
2	0	0	156	4	0	0	0	0	0	0
3	0	0	3	157	0	0	0	0	0	0
4	2	0	0	0	158	0	0	0	0	0
5	0	0	0	0	0	160	0	0	0	0
6	0	0	0	1	0	0	158	0	1	0
7	0	0	0	0	0	0	0	160	0	0
8	0	0	0	0	0	2	0	0	158	0
9	1	1	0	1	0	0	0	0	0	157

Tabla 5.2. Matriz de confusión para palabras aisladas en inglés de la base de TI, método DTW, con parámetros KLT, trayectorias tipo I, precisión 98 4%

Este método se aplicó también a nuestra base de datos en español grabada en ruido ambiente, calculado en aproximadamente 52 dB, lo que resulta en un SNR de alrededor de 25 dB. Se utilizaron 5 hombres y 5 mujeres, para un total de 1 400 repeticiones para entrenamiento y el mismo número para reconocimiento Con las trayectorias modificadas, la precisión fue de 98.64%, que es muy alta para las condiciones de ruido, ver tabla 5.3. En los errores destacan entre *dos* y *uno*, y entre *seis* y *tres*. En el primer caso se observa en los errores, una fricativa muy débil /t/. para el segundo caso se observa que los tres últimos

segmentos son muy parecidos, y no existe una marcada diferencia en las palabras mal reconocidas, entre el fonema /e/ y la secuencia /e/-/i/.

	0	1	2	3	4	5	6	7	8	9
0	139	0	0	1	0	0	0	0	0	0
1	1	138	0	0	1	0	0	0	10	0
2	1	3	136	0	0	0	0	0	0	0
3	1	0	0	137	0	0	2	0	0	0
4	0	0	0	0	140	0	0	0	0	0
5	0	0	0	0	0	140	0	0	0	0
6	0	0	0	3	0	0	137	0	0	0
7	1	0	0	1	0	0	0	138	0	0
8	0	0	1	0	2	0	0	0	137	0
9	0	0	0	1	0	0	0	0	0	139

Tabla 5.3. Matriz de confusión para palabras aisladas en español con alto ruido ambiental, método DTW con parámetros KLT, trayectorias modificadas, precisión 98.64%.

Estos experimentos se compararon con el mismo sistema DTW, para los parámetros de análisis LPC y cepstral, usando para LPC la distancia de Itakura-Saito, con la base de datos de TI. La precisión de KL es un poco menor a estos dos. Los resultados para LPC se muestran en la tabla 5.4.

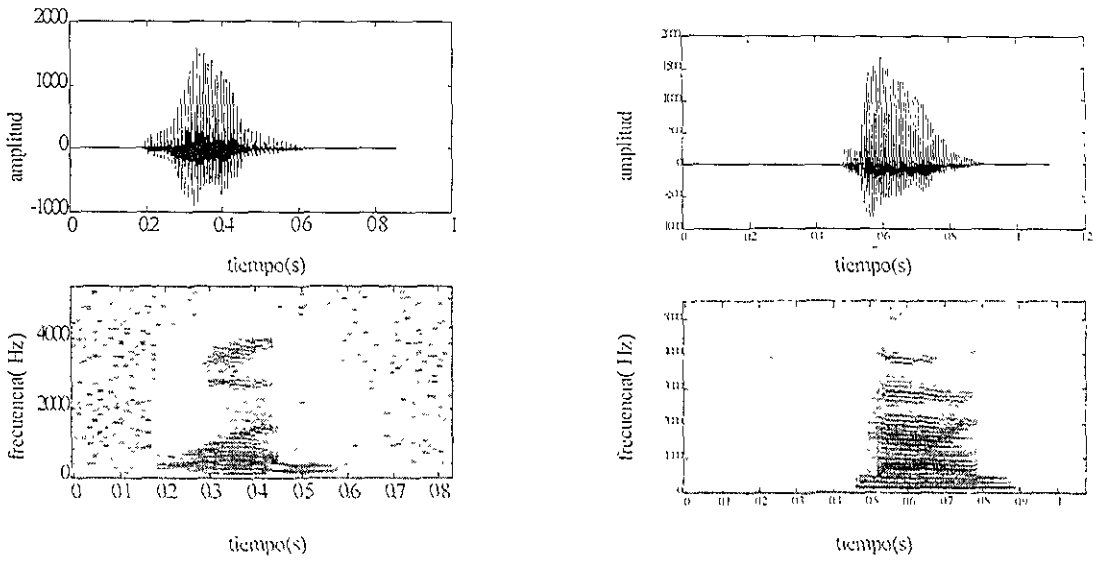
	0	1	2	3	4	5	6	7	8	9
0	160	0	0	0	0	0	0	0	0	0
1	0	158	0	0	0	0	0	0	0	2
2	0	0	160	0	0	0	0	0	0	0
3	1	0	0	159	0	0	0	0	0	0
4	1	0	0	0	159	0	0	0	0	0
5	0	0	0	0	0	160	0	0	0	0
6	0	0	0	0	0	1	159	0	0	0
7	0	0	0	0	0	0	0	160	0	0
8	0	0	0	0	0	0	0	0	159	0
9	0	1	0	0	0	0	0	0	0	159

Tabla 5.4. Matriz de confusión para palabras aisladas en inglés de la base de TI, método DTW con parámetros LPC, trayectorias modificadas, precisión 99.56%.

Para los parámetros cepstral se obtuvieron rendimientos un poco mejores que ambos, se utilizó la distancia euclidiana, ver tabla 5.5. Sin embargo, cabe hacer notar que el tiempo de procesamiento en la clasificación es de aproximadamente el doble, 90 s/palabra, comparado con KL, de 51 s/palabra. También es importante mencionar que acerca de los patrones de comportamiento en los errores, en el caso de "one" y "nine" en el idioma inglés, cuya similitud es evidente. Otros casos como "three" o "four" con "zero", se observó una repetición contaminada con ruidos del micrófono, y de *pops*.

procesamiento en la clasificación es de aproximadamente el doble, 90 s/palabra, comparado con KL, de 51 s/palabra.

Es importante mencionar acerca de los patrones de comportamiento en los errores, que estos se produjeron en el caso de "one" y "nine" en el idioma inglés; se observan los errores cuando la primera nasal de "nine" es débil y los fonemas vocálicos de ambas palabras se asemejan, por la propia dicción del hablante. Estos efectos se muestran en la figura 5.6, donde se muestra una repetición mal clasificada de "one" y una repetición de entrenamiento de "nine" cercana a la anterior, se presentan tanto las señales en el tiempo como sus espectrogramas. En la repetición de "one" mostrada, el fonema vocálico difiere sustancialmente de otras repeticiones de esta palabra con otros hablantes , sin embargo, en otras repeticiones de "nine", este efecto se ve compensado por su primera nasal, lo que en esta repetición de no se da y se provoca el error.



(a) (b)

Figura 5 6. Señales en el tiempo y espectogramas de (a) una repetición mal clasificada de "one" y (b)una repetición cercana a ésta de "nine".

Otros casos como "three" o "four" con "zero", se observó una repetición contaminada con ruidos del micrófono, y de *pops*, que es también el caso de los errores con la base de datos en español.

5.2.Clasificación usando VQ multiseccionada

5.2.1.Cuantización vectorial

La cuantización vectorial (VQ) es utilizada en procesamiento de voz no para generar nuevas unidades por trama como los son los coeficientes LPC o cepstral, sino para comprimir los vectores por trama durante la etapa de entrenamiento .

La cuantización vectorial puede interpretarse como una forma de reconocimiento de patrones donde un vector de entrada es aproximado por un patrón dentro de un conjunto predeterminado de estos, o en otro lenguaje, el vector de entrada es transformado en un patrón dentro de un conjunto finito de patrones llamados centroides. Así VQ va más allá de una mera generalización de cuantización escalar. Su alcance e implicaciones son vastas, entre las que destacan el reconocimiento de voz y la compresión de voz e imágenes. Se describirán algunos elementos básicos de la teoría de VQ[73].

Un vector de cuantización Q de dimensión k y tamaño N es un mapeo de un vector en el espacio k dimensional, R^k, a un conjunto finito C conteniendo N salidas o puntos de reproducción, llamado código de vectores o conjunto de centroides, $Q: R^k \rightarrow C$.

Asociado a C existe una partición de R^k en N regiones o celdas, R_i para $i \in J$. La i^{th} región esta definida por $R_i = \{x \in R^k : Q(x) = y_i\}$, donde

$$\bigcup_i R_i = R^k \; y \quad R_i \cap R_j = \varnothing \; para \quad i \neq j, \qquad \text{Ec 5. 12}$$

una propiedad importante de un conjunto de R^k es la convexidad. Un cuantizador de vector es llamado regular si (a) cada celda, R_i, es un conjunto convexo, y (b) si $x \in R_i$, entonces $Q(x)(= y_i)$ está contenido en R_i. Un cuantizador es politopal si es regular y sus celdas de partición son limitadas por segmentos de superficies hiperplanas en *k* dimensiones. Un cuantizador regular en el caso de una dimensión es siempre politopal. Un cuantizador es delimitado cuando no tiene ninguna región de traslape.

Una clase especial importante de cuantizadores son los llamados de Voronoi o de vecindad cercana, tienen la característica de que la partición está completamente determinada mediante el libro de códigos y una medición de la distorsión, sus regiones son tales que

$$R_i = \{x : d(x, y_i) \le d(x, y_j), \quad \forall j \in J\} \qquad \text{Ec 5. 13}$$

una ventaja de esta clase de codificadores es que el proceso de codificado no requiere de ningún almacenamiento explícito de la descripción geométrica de las celdas, estos cuantizadores son óptimos.

Se han diseñado diversas técnicas para la obtención de dentroides que son óptimas y conducen a regiones de Voronoi. Entre estas técnicas destacan las de k-medias, LBG (Linde-Buzo-Gray) e ISODATA (Iterative Self-Organizing Data Analysis Techniques A) . Se describirán brevemente.

Algoritmo K medias.

Dada la función de criterio:

$$J_e = \sum_{j=1}^{K} \sum_{x \in \chi_j} \left\| x - z_j \right\|^2 \qquad \text{Ec. 5 14}$$

donde: K es el número de agrupaciones, z_j es el centroide para la región j y χ_j es el conjunto de muestras asignadas a la agrupación j.

Los pasos del algoritmo son[74].

1. Seleccionar K centroides iniciales $z_1(1), z_2(1), \ldots, z_K(1)$.
2. En la iteración l, asignar muestras a las agrupaciones, esto es, asignar x a $\chi_i(l)$ si
$$\| x - z_i(l) \| \le \| x - z_j(l) \|, \quad j = 1,2,\ldots K \qquad i \ne j \qquad \text{Ec. 5.15}$$
3. Calcular nuevos centroides:
$$z_i(l+1) = \frac{1}{N_i} \sum_{x \in \chi_i(l)} x \qquad i = 1,2,\ldots K \qquad \text{Ec 5.16}$$

donde N_i es el número de muestras en $\chi_i(l)$

4. Si $z_i(l+1) = z_i(l)\ para\quad i = 1,2,\ldots,K$, el algoritmo converge y termina. De otra forma, regresar al paso (2).

El comportamiento es influido por el número de centroides escogidos (K), la selección de centroides iniciales, el orden en el cual las muestras son tomadas y las propiedades geométricas de los datos. Es deseable que los resultados se comparen con distintas variaciones de K y los centroides iniciales.

Método LBG (Linde-Buzo-Gray)

Consta de los siguientes pasos después de fijar una distancia de distorsión D_T y un valor de separación ρ [75]:

1. Escoger un centroide inicial z_1,
2. Se calcula el centroide o centroides de acuerdo con $z_{1i} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$ para las muestras x_i de dimensión n.
3. Si $d_i > D_T$ entonces se separa z_1 en los centroides $\begin{aligned} z_1' &= z_1 + (\rho,\ldots,\rho) \\ z_2' &= z_1 - (\rho,\ldots,\rho) \end{aligned}$
   Ir al paso 4. En caso contrario el algoritmo converge.
4. Se clasifican las muestras x_i en χ_i e ir a paso 2.

Observe que en este método se da prioridad a que el agrupamiento no sobrepase cierta distorsión D_T.

ISODATA (Iterative Self-Organizing Data Analysis Techniques A)

Los pasos del algoritmo son los siguientes[74][76]:

1. Especificar los parámetros: K_{max} es el número máximo de agrupamientos deseados, N_{min} es el número mínimo de muestras en un agrupamiento, S_{max} es la desviación estándar máxima de un agrupamiento, D_{min} es la distancia mínima entre agrupamientos, I_{max} es el número máximo de iteraciones, y z_j es el centroide inicial del agrupamiento, 1=1,..., K
2. Asignar muestras a regiones, esto es asignar x a χ_i si

$$\|x - z_i\| < \|x - z_j\|, \quad i = 1,2,\ldots,K, \quad j \neq i \qquad \text{Ec. 5.17}$$

3. Eliminar regiones con muestras menores a N_{min}

Si para cualquier i, $N_i < N_{min}$, se borrará z_i del conjunto de centroides y se decrementará K: $K = K - 1$, donde N_i = número de muestras en la región χ_i.

4. Cálculo de estadísticas del centroide,

a) Actualizar centorides de grupo

$$z_i = \frac{1}{N_i} \sum_{x \in \chi_i} x \qquad i = 1,2,\ldots,K \qquad \text{Ec 5.18}$$

b) Calcular la distancia promedio $\bar{d}_i$ de las muestras en el grupo χ_i para el centroide del grupo z_i

$$\bar{d}_i = \frac{1}{N_i} \sum_{x \in \chi_i} \|x - z_i\| \qquad i = 1,2,\ldots,K \qquad \text{Ec. 5.19}$$

c) Calcular el promedio de todas las distancias de muestras para los centroides del grupo donde N es el número total de muestras

$$\bar{d} = \frac{1}{N} \sum_{i=1}^{K} N_i \bar{d}_i \qquad \text{Ec. 5.20}$$

d) Incrementar la iteración del índice: $I = I + 1$.

5. Decidir si los centroides se dividen o se combinan,

Si $I = I_{max}$, ir al paso 7 (combinar-se termina)

o, si $k \leq \frac{1}{2} K_{max}$, ir al paso 6 (dividir)

o, si I es par o $K \geq 2K_{max}$, ir al paso 7 (combinar)

de otro modo, ir al paso 6 (dividir).

6 Dividir agrupamientos

a) Calcular el vector de desviación estándar del grupo

$\sigma_i = [\sigma_{i1}, \sigma_{i2}, \ldots, \sigma_{in}]^T, \quad i = 1,2,\ldots,K$

con

$$\sigma_{ik}^2 = \frac{1}{N_i} \sum_{x \in \chi_i} (x_k - z_{ik})^2 \qquad \text{Ec. 5.21}$$

donde n es la dimensión de las muestras, x_k es la k componente en la región χ_i y z_{ik} es la k componente de z_i.

b) Encontrar la componente máxima de cada σ_i, esto es, $\sigma_{i_{max}} = \max_k \sigma_{ik}, \quad i = 1,2,\ldots,K$

c) Dividir los grupos con desviación estándar grande

$$\text{Si } \sigma_{i_{max}} > s_{max} \quad \text{y} \quad (1)\, K \le \frac{1}{2} K_{max} \quad \text{o} \quad (1)\, \bar{d}_i > \bar{d}$$

Entonces dividir z_i en z_i^+ y z_i^-

$$z_i^+ = z_i + [0,\ldots,0,\sigma_{i_{max}},0,\ldots,0]^T$$
$$z_i^- = z_i - [0,\ldots,0,\sigma_{i_{max}},0,\ldots,0]^T \qquad \text{Ec.5.22}$$

donde los valores diferentes de cero son puestos en el componente máximo del vector de desviación estándar.

Ir al paso 2 (nueva iteración), de otra forma ir al paso 7 (combinar)

7. Combinar los grupos,

a) Calcular las distancias entre grupos

$$d_{ij} = \|z_i - z_j\|, \quad i = 1,2,\ldots,K-1, \quad j = i+1,\ldots,K$$

b) Combinar los grupos con pequeñas distancias entre grupos

Encontrar las distancia d_{ij} tales que

$$d_{ij} < d_{min}$$

Arreglar en orden ascendente

$$[d_{i_1 j_1}, d_{i_2 j_2}, \ldots, d_{i_L j_L}]$$

Considere la distancia restante más pequeña. $d_{i_1 i_2}$, con un centroide de grupo asociado z_{i_1} y z_{j_1}. Si ningún grupo ha sido combinado en esta iteración, hay que combinarlo formando un nuevo centroide de grupo

$$z_I = \frac{1}{N_{i_1} + N_{i_2}} \left[N_{i_1} Z_{i_1} + N_{i_2} z_{i_2} \right] \qquad \text{Ec. 1.23}$$

borrar z_{i_1} y z_{i_2} y decrementar K
Repetir para las siguientes distancias más pequeñas.

8. Terminar o continuar:

Si $I = I_{max}$, terminar, de otro modo, ir al paso 2.

Este método es el más completo. Sin embargo, varios pasos pueden omitirse dependiendo del tipo de datos. De hecho, los dos primeros métodos mencionados deben complementarse con estrategias de ISODATA de acuerdo con el problema en particular.

5.2.2.Método de clasificación

En el modo de entrenamiento, ver figura 5.6, el sistema genera 16 centroides por cada vector propio en cada segmento. El número de segmentos se debe fijar para cada prueba, los

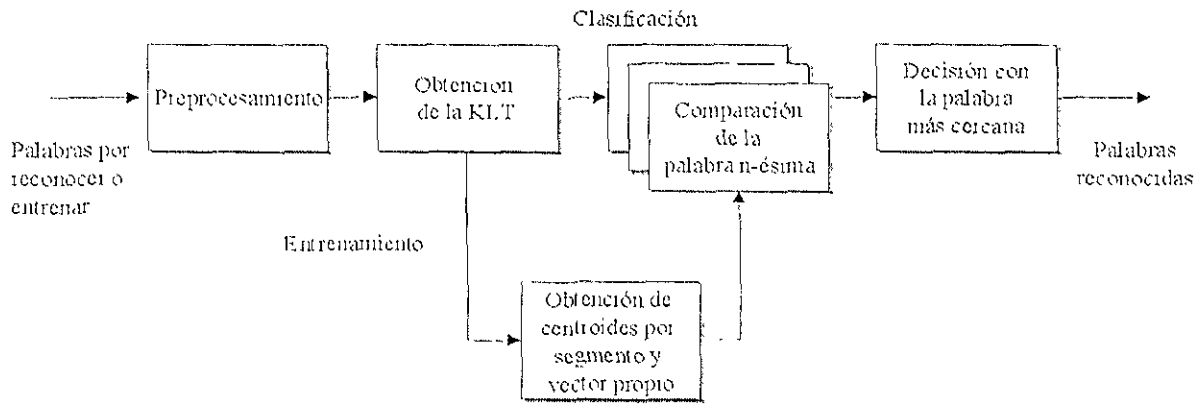


Figura 5.7. Sistema de reconocimiento de palabras aisladas.

segmentos utilizados fueron de 4 a 8 segmentos. el algoritmo de agrupamiento es de K-medias modificada al introducir pasos 3, 5, y 7 adaptados del método ISODATA, mencionados en el tema 1.2.3, esta técnica fue también usada en la clasificación.

En la clasificación, para cada palabra sus subpalabras son comparadas con los centroides de las palabras de referencia, la información requerida por subpalabra son los vectores y valores propios. Entonces la palabra se clasifica con el mínimo de las distancias por subpalabra, usando la distancia basada en Hilbert-Schmidt, mencionada en el capítulo 4, ecuación 4.39.

	0	1	2	3	4	5	6	7	8	9
0	160	0	0	0	0	0	0	0	0	0
1	0	157	0	0	0	0	0	0	0	2
2	0	0	160	0	0	0	0	0	0	0
3	0	0	0	160	0	0	0	0	0	0
4	1	0	0	0	160	0	0	0	0	0
5	0	0	0	0	0	159	0	0	0	1
6	0	0	0	0	0	0	160	0	0	0
7	0	0	0	0	0	0	2	157	0	0
8	0	0	0	0	0	1	3	0	157	0
9	0	3	0	0	0	0	0	0	0	156

Tabla 5.6. Matriz de confusión para palabras aisladas en inglés de la base de TI, usando VQ de 4 segmentos con parámetros KLT, precisión 99 24%.

	0	1	2	3	4	5	6	7	8	9
0	160	0	0	0	0	0	0	0	0	0
1	0	156	0	0	0	0	0	1	0	3
2	0	0	156	4	0	0	0	0	0	0
3	0	0	0	160	0	0	0	0	0	0
4	0	0	0	0	159	1	0	0	0	0
5	0	0	0	0	0	159	0	0	0	1
6	0	0	0	0	0	0	160	0	0	0
7	0	0	0	0	0	0	0	160	0	0
8	0	0	0	2	0	0	0	0	157	0
9	0	1	0	0	0	1	0	0	0	158

Tabla 5.7. Matriz de confusión para palabras aisladas en inglés de la base de TI, usando VQ de 4 segmentos, con parámetros LPC, , precisión 99.06%.

Para la base de datos en inglés se utilizaron 1 000 repeticiones con 10 repeticiones de cada dígito. La fase de clasificación es dependiente del locutor, se usaron 1 600 repeticiones , con 16 repeticiones de cada dígito. La precisión más alta obtenida fue de 99.24%, ver tabla 5.6, las otras tuvieron precisiones mayores a 95%. El mayor problema se tuvo a partir de 7 segmentos, ya que los vectores para agrupar disminuían en su número por lo que los centroides no eran ya muy representativos.

Estos experimentos se compararon con el mismo tipo de VQ, para los parámetros de análisis LPC y cepstral, en el primero se usó la distancia de Itakura-Saito. La precisión de KL es un poco mejor que ambas LPC. En la tabla 5.7 se muestran los resultados para LPC con 4 segmentos MLR, se tiene una precisión de 99.06%. Las precisiones se mantienen en este nivel para otros segmentos, pero son menores.

Para los parámetros cepstral con 4 segmentos se obtuvo un rendimiento de 99.12%, como se observa en la tabla 5.8. La precisión tuvo un máximo para 5 segmentos de 99.40%. Se utiliza la distancia euclidiana con estos parámetros.

	0	1	2	3	4	5	6	7	8	9
0	160	0	0	0	0	0	0	0	0	0
1	0	157	0	0	0	0	0	0	0	3
2	0	0	158	1	0	0	0	0	0	0
3	0	0	1	159	0	0	0	0	0	0
4	0	0	0	0	160	0	0	0	0	0
5	0	0	0	0	0	158	0	0	0	1
6	0	0	0	0	0	0	160	0	0	0
7	0	0	0	0	0	0	2	158	0	0
8	0	0	0	1	0	0	0	0	159	0
9	0	1	0	0	0	1	0	0	0	157

Tabla 5.8. Matriz de confusión para palabras aisladas en inglés de la base de TI, usando VQ de 4 segmentos, con parámetros cesptral, precisión 99 12%

	0	1	2	3	4	5	6	7	8	9
0	48	0	0	0	0	1	1	3	0	2
1	0	55	0	0	0	0	0	0	0	0
2	0	0	55	0	0	0	0	0	0	0
3	0	0	0	48	0	0	6	0	0	1
4	0	0	0	0	53	0	0	0	2	0
5	1	0	0	0	0	54	0	0	0	0
6	0	0	0	2	0	0	50	3	0	0
7	2	0	1	2	0	1	3	46	0	0
8	0	0	0	0	3	0	0	0	52	0
9	0	0	0	0	0	0	0	0	0	55

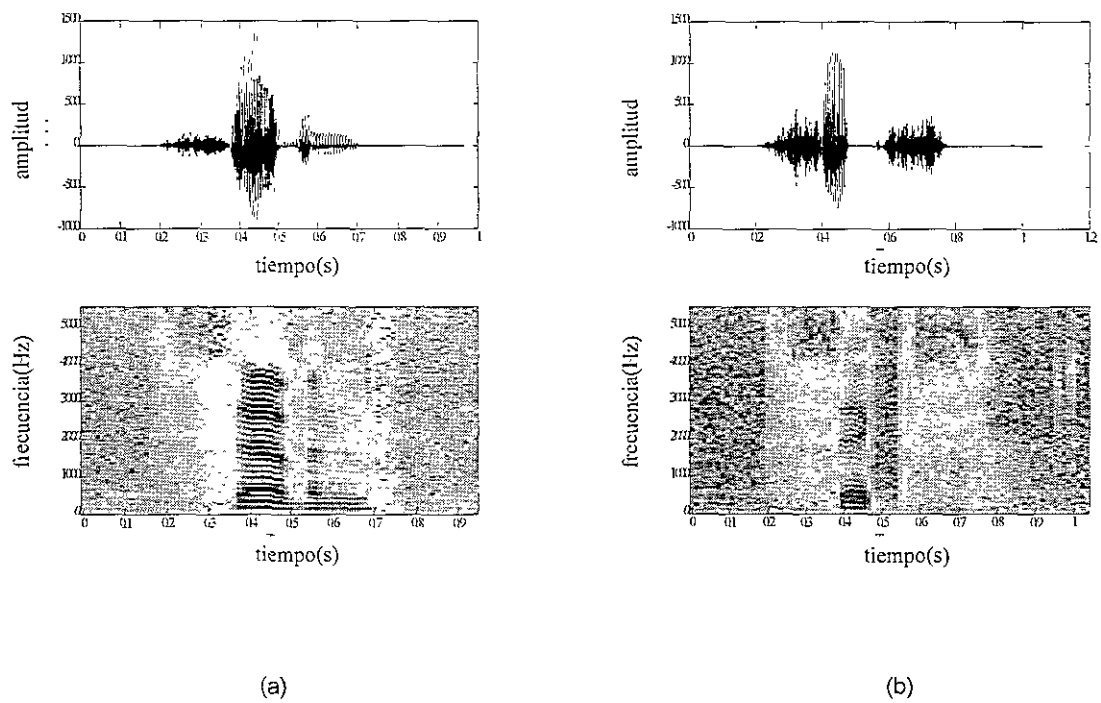
Tabla 5.9 Matriz de confusión para palabras aisladas en español con alto ruido ambiental, usando VQ de 4 segmentos, con parámetros KLT, precisión 93 82%.

Para la base de datos en español se utilizaron 400 repeticiones de los dígitos para entrenamiento y 550 para clasificación En este caso el ruido ambiental influyó notablemente en el comportamiento de los centroides KLT, aun usando 32 centroides por segmento no se logró separar las clases En consecuencia el rendimiento fue de 93.82%, ver tabla 5.9.

El rendimiento señalado para KLT en español no disminuye de esta forma para parámetros LPC o cepstral, el mejor es para cepstrals , como se muestra en la tabla 5.10.

	0	1	2	3	4	5	6	7	8	9
0	52	0	0	0	0	1	0	0	0	2
1	0	54	0	0	1	0	0	0	0	0
2	0	0	55	0	0	0	0	0	0	0
3	0	0	0	54	0	0	0	0	0	1
4	0	0	0	0	53	0	0	0	2	0
5	0	0	0	0	0	54	0	1	0	0
6	0	0	0	1	0	0	54	0	0	0
7	0	0	0	0	0	0	0	55	0	0
8	0	0	0	0	2	0	0	0	53	0
9	0	0	0	0	0	0	0	0	0	55

Tabla 5.10. Matriz de confusión para palabras aisladas en español con alto ruido ambiental, usando VQ de 4 segmentos, con parámetros cesptral, precisión 99.12%.



(a) (b)

Figura 5.8. Señales en el tiempo y espectogramas de (a) una repetición mal clasificada de "seven" y (b)una repetición cercana a ésta de "six".

De la revisión realizada al comportamiento del VQ con parámetros KLT, se pudo observar que el problema recae en la baja discriminación entre centroides y vectores de entrada que corresponden a fonemas muy distintos, e inclusive al cruce de centroides de fonemas semejantes, como algunos vocálicos. Este problema se podría subsanar aumentando el

número de subpalabras, sin embargo, se tendría como contraparte un mayor tiempo de procesamiento.

En particular, se observan en las tablas 5.6 a 5.8 patrones de errores con las palabras "one" y "nine", y aparecen más errores entre "six" y "seven". En este último caso, se muestra un ejemplo en las figuras 5.8 y 5.9.

En la figura 5.8 se muestra la señal en el tiempo y el espectrograma de repeticiones de las palabras "seven" y "six", para dos hablantes diferentes. La mayoria de los errores corresponden a la palabra "seven" con el centroide de la palabra "six". Esto sucedió más para un hablante en específico, donde en la palabra "seven": la oclusiva /b/ tomó un alófono fricativo, /, el segundo fonema vocálico es débil, y la oclusiva es dominante sobre la nasal /n/, como se observa en la figura 5.8.

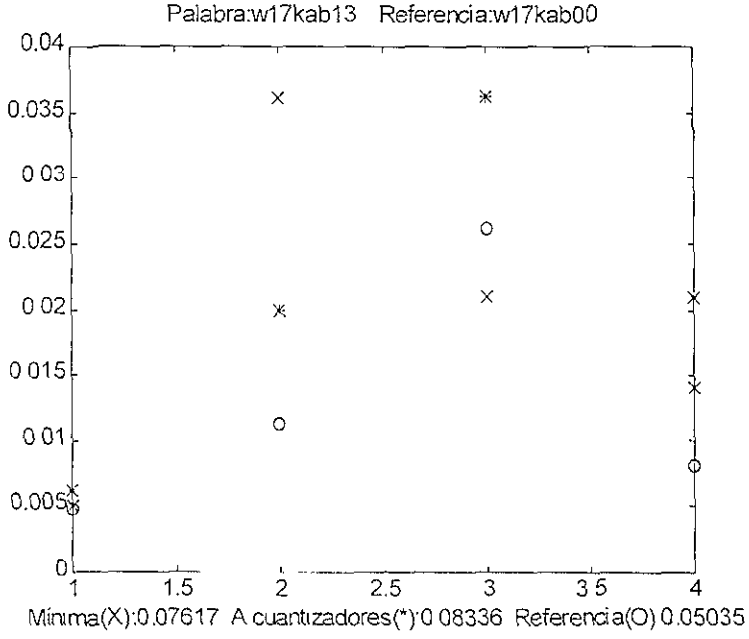


Figura 5 9 Distancias en 4 subpalabras de una repetición de "seven" con los centroides de "seven", "six", y la repetición mas próxima de "seven". Clasificación usando HMM

En la figura 5.9 se muestran las distancias para cada una de las 4 subpalabras de una repetición de "seven" mal clasificada como "six". Con el símbolo (X) se señalan las distancias con los centroides de "six", con (*) se muestran las distancias a los centroides de "six", y con (O) las distancias a la repetición de "seven" más cercana a ésta De esta figura pueden observarse como las distancias que producen la mal decisión corresponden a las conclusiones

señaladas en el párrafo anterior. Este patrón de comportamiento se produce con todos las repeticiones de "seven" mal clasificadas, aún con distintas segmentaciones, y los errores se producen por el mismo hablante. En la figura 5.10, se observa como las distancias entre la repetición y los centroides tienen el mismo patrón que en la figura 5.9 para 4 segmentaciones.

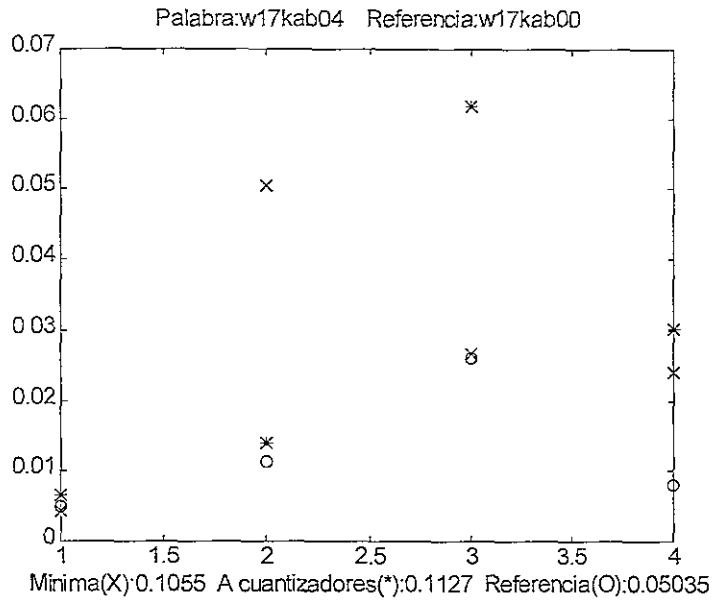


Figura 5.10. Distancias en 4 subpalabras de una repetición de "seven" con los centroides de "seven", "six", y la repetición más próxima de "seven". Clasificación usando HMM

5.3.Clasificación usando HMM

5.3.1. Procesos de Markov

Una cadena o proceso discreto de Markov es un modelo donde una variable aleatoria puede tener ciertos valores llamados estados $S_1, S_2, \ldots, S_N$ y la variable cambia de estado o pertenece a un estado de acuerdo con cierta probabilidad El estado presente de la variable se denota q_t y depende en general de los estados precedentes.

Un proceso discreto de Markov de primer orden depende solamente del estado precedente más reciente, esto es:

$$P[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \ldots] = P[q_t = S_j | P[q_{t-1} = S_i] \qquad \text{Ec. 5. 24}$$

Aún más, considerando solamente los procesos en los cuales el lado derecho de la ecuación 5.9 es independiente del tiempo, entonces se da lugar a las llamadas probabilidades de transición a_{ij} definidas como

$$a_{ij} = P[q_t = S_j | q_{t-1} = S_i], \quad 1 \le i, j \le N \qquad \text{Ec. 5. 25}$$

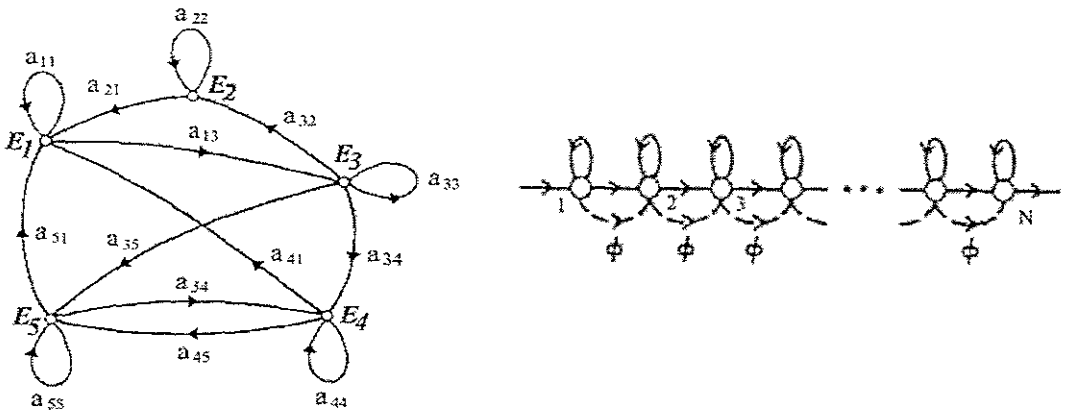
y tienen las propiedades $a_{ij} \ge 0$ y $\sum_{j=1}^{N} a_{ij} = 1$

La matriz A de probabilidades de transición de estados se define por $A = (a_{ij})$. Las probabilidades de estados iniciales π_i se definen por

$$\pi_i = P[q_1 = S_i] \quad 1 \le i \le N \qquad \text{Ec. 5. 26}$$

En la figura 5 11 se ilustran dos modelos clásicos de Markov El proceso mostrado en (a) es ergódico o totalmente interconectado, en el cual cada estado puede ser alcanzado por cualquier otro. Para aplicaciones en voz son más útiles los modelos de izquierda a derecha,

inciso (b), en el cual las probabilidades de transición tienen la propiedad $a_{ij} = 0$, cuando $j < i$ y las probabilidades de estados iniciales tienen la propiedad $\pi_1 = 1$ y $\pi_i = 0$ si $i \neq 1$, y frecuentemente se considera la limitante adicional $a_{ij} = 0$ para $j > i + \Delta$.



(a) (b)

Figura 5.11. Rutas de ajuste en el tiempo, la óptima se señala por la línea sombreada.

En algunos casos el estado no corresponde a un evento físico u observable, así que se extiende el concepto de los procesos de Markov a otro donde las observaciones son variables aleatorias en los estados. A este modelo se le llama modelo oculto de Markov, *HMM* por sus siglas en inglés, que se establece entonces como un doble proceso estocástico, uno de estos no observable, donde éste solamente puede ser descrito a través del otro proceso estocástico que produce una secuencia de observaciones. Así en un tiempo t el estado S_j es seleccionado aleatoriamente, esto es $q_t = S_j$, con probabilidad a_{ij} donde i representa el estado previo. Ya dentro de este estado S_j el símbolo observado O_k es seleccionado con probabilidad $b_j(k)$, donde $\sum_{k=1}^{M} b_j(k) = 1$

Los elementos de un modelo discreto oculto de Markov son entonces:

1. El número N de estados.

2. El número M de símbolos de observaciones por estado. Las observaciones corresponden a las salidas físicas del sistema que está siendo modelado. Denotamos al conjunto de símbolos como $V = \{v_1, v_2, \ldots, v_M\}$.

3. La matriz de transiciones $A = (a_{ij})$, donde

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i] \qquad 1 \le i,\ j \le N \qquad \text{Ec 5. 27}$$

4. La matriz de probabilidades de los símbolos observados $B = (b_j(k))$ donde

$$b_j(k) = P[v_k \ en\ t | q_t = S_j], \qquad 1 \le j \le N,\ 1 \le k \le M \qquad \text{Ec. 5. 28}$$

5. El vector de estados iniciales $\Pi = \{\pi_i\}$, donde

$$\pi_i = P[q_1 = S_i], \qquad 1 \le i \le N \qquad \text{Ec. 5. 29}$$

El modelo oculto de Markov (HMM) $\lambda = (A, B, \Pi)$ puede ser usado para generar una secuencia de observaciones o describir como una secuencia dada fue generada por un modelo HMM. Estos son precisamente los dos problemas básicos para el modelo, esto es:

1. Dada la secuencia de observaciones $O = O_1, O_2, \ldots, O_t$ como ajustamos los parámetros del modelo A, B y Π para optimar el modelo λ, de manera tal que describa de la mejor forma la secuencia dada. Las secuencias utilizadas para ajustar el modelo son llamadas secuencias de entrenamiento.

2. Dada una secuencia de observaciones O y un modelo λ cómo calculamos eficientemente $P(O|\lambda)$, y cómo buscamos una secuencia del modelo que maximice la probabilidad para una secuencia dada.

Para aclarar, consideremos el siguiente sistema de reconocimiento de palabras aisladas. Para cada palabra W del vocabulario, diseñamos un modelo HMM de N estados. Representemos la señal de voz de una palabra como una secuencia de vectores espectrales codificados, donde la codificación se hace usando un libro de códigos con M vectores

espectrales, entonces cada observación es el índice del vector espectral más cercano. Así, para cada palabra del vocabulario tenemos secuencias de entrenamiento que consisten en varias repeticiones de la palabra, convertidas en secuencias de índices del código. La primera tarea es optimar los parámetros del modelo por palabra usando las secuencias de entrenamiento. La labor de reconocimiento de una nueva repetición es el segundo problema, y consiste en evaluar las probabilidades $P(O|\lambda)$ para cada modelo de palabra y escoger el de probabilidad más alta.

Estos dos problemas están ligados, y se empezará por el segundo.

Solución del problema 2

Dado un modelo λ y una secuencia de observaciones $O = O_1, O_2, \ldots, O_T$, para calcular $P(O|\lambda)$ tenemos que enumerar todas las posibles secuencias de estados de longitud T. Considere una secuencia fija de estados $Q = q_1 q_2 \ldots q_T$

La probabilidad de la secuencia de observaciones O para esta secuencia de estados es

$$P(O \mid Q, \lambda) = \prod_{t=1}^{T} P(O_t \mid q_t, \lambda) = b_{q_1}(O_1) b_{q_2}(O_2) \ldots b_{q_T}(O_T) \qquad \text{Ec. 5. 30}$$

donde asumimos independencia estadística de las observaciones.

La probabilidad de la secuencia de estados Q es:

$$P(Q \mid \lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \ldots a_{q_{T-1} q_T} \qquad \text{Ec. 5. 31}$$

La probabilidad conjunta de O y Q es el producto de las ecuaciones 5.15 y 5.16:

$$P(O, Q \mid \lambda) = P(O \mid Q, \lambda) P(Q \mid \lambda) \qquad \text{Ec. 5. 32}$$

La probabilidad de O dado el modelo se obtiene sumando esta probabilidad conjunta sobre todas las posibles secuencias de estados q, es decir:

$$P(O,Q|\lambda)=\sum_{todaQ}P(O|Q,\lambda)P(Q,\lambda)=\sum_{q_1,q_2,\ldots,q_T}\pi_{q_1}b_{q_1}(O_1)a_{q_1q_2}b_{q_2}(O_2)\ldots a_{q_{T-1}q_T}b_{q_T}(O_T) \quad \text{Ec. 5.33}$$

Los cálculos para $P(O|\lambda)$, según la definición anterior, involucran $(2T-1)N^T$ multiplicaciones y N_1^T adiciones, esto es del orden de $2TN^T$ cálculos. Esto es demasiado laborioso aun para valores pequeños de N y de T; por ejemplo para $N=5$ estados y $T=50$ observaciones se tienen del orden de $100(5^{50})$ cálculos. Un procedimiento más eficiente existe y es llamado procedimiento *hacia adelante-hacia atrás*. Este algoritmo solamente reordena los cálculos involucrados en la definición de $P(O|\lambda)$ con el fin de reducirlos. Se definen las variables

$$\alpha_t(i)=P(O_1O_2\ldots O_t,q_t=S_i|\lambda) \quad \text{Ec. 5. 34}$$

$$\beta_t(i)=P(O_{t+1}O_{t+2}\ldots O_T|q_t=S_i,\lambda) \quad \text{Ec. 5. 35}$$

llamadas variables hacia adelante y hacia atrás. La probabilidad $P(O|\lambda)$ está dada por cualquiera de las dos siguientes expresiones:

$$P(O|\lambda)=\sum_{i=1}^{N}\alpha_T(i)=\sum_{i=1}^{N}\alpha_t(i)\beta_t(i) \quad \text{Ec. 5. 36}$$

El cálculo de $\alpha_t(i)$ y $\beta_t(i)$ se puede hacer mediante las fórmulas inductivas:

$$\alpha_1(i)=\pi_i b_i(O_1), \quad 1\le i\le N \quad \text{Ec. 5. 37}$$

$$\alpha_{t+1}(j)=[\sum_{i=1}^{N}\alpha_t(i)a_{ij}]b_j(O_{t+1}), \quad 1\le t\le T-1,\ 1\le j\le N \quad \text{Ec. 5. 38}$$

$$\beta_T(i)=1 \quad 1\le i\le N \quad \text{Ec. 5. 39}$$

$$\beta_t(i)=\sum_{j=1}^{N} a_{ij} b_j(O_{t+1})\beta_{t+1}(j), \quad t=T-1,T-2,\ldots,1 \quad 1\le i\le N \qquad \text{Ec. 5. 40}$$

En lugar de calcular la probabilidad $P(O|\lambda)$, que implica los cálculos sobre todas las posibles secuencias de estados, podemos encontrar una secuencia óptima del modelo asociada a la secuencia de observación dada y obtener la probabilidad para esta secuencia. Existen diferentes criterios de optimación. Una técnica para encontrar la mejor secuencia está basada en programación dinámica y es llamada Algoritmo de Viterbi.

Para diseñar este algoritmo, necesitamos definir la variable

$$\gamma_t(i)=P(q_t=S_i|O,\lambda)=\frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} \qquad \text{Ec. 5. 41}$$

esto es, la probabilidad de estar en el estado S_i en un tiempo t, dada la secuencia de observaciones O y el modelo λ. Usando γ podemos obtener el estado más adecuado q_t, en el tiempo t como:

$$q_t=\arg\max_{1\le i\le N}[\gamma_t(i)], \quad 1\le t\le T \qquad \text{Ec. 5. 42}$$

El mejor puntaje a lo largo de una sola ruta, en un tiempo t, el cual lleva el conteo para las primeras t observaciones y termina en el estado S_i es:

$$\delta_t(i)=\max_{q_1,q_2,\ldots,q_{t-1}} P[q_1q_2\ldots q_t=S_i,O_1O_2\ldots O_t|\lambda] \qquad 5.43$$

El procedimiento completo del algoritmo Viterbi es el siguiente:

Inicia:

$$\delta_1(i)=\pi_i b_i(O_1), \quad 1\le i\le N \qquad \text{Ec. 5. 44}$$

$$\psi_1(i)=0 \qquad \text{Ec. 5. 45}$$

Obtiene variables de Viterbi: Para $t \geq 2, \quad 1 \leq j \leq N$

$$\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] \, b_j(O_t), \qquad \text{Ec. 5. 46}$$

$$_t(j) = \arg \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}], \qquad \text{Ec. 5. 47}$$

Calcula la probabilidad:

$$\delta^* = \max_{1 \leq i \leq N} [\delta_t(i)], \qquad \text{Ec. 5. 48}$$

$$i_t^* = \max_{1 \leq i \leq N} [\delta_t(i)], \qquad \text{Ec. 5. 49}$$

Obtiene secuencia: Para $t = T-1, T-2, \ldots, 1$

$$i_t^* = \psi_{t+1}(i_{t+1}^*), \qquad \text{Ec. 5. 50}$$

Solución del problema 1

Este problema trata de ajustar los parámetros del modelo (A, B, Π) para maximizar la probabilidad de una secuencia de observaciones dado el modelo. No existe una manera analítica para solucionar el modelo que maximice la probabilidad de la secuencia de observaciones. Sin embargo, podemos escoger $\lambda = (A, B, \Pi)$ tal que $P(O \mid \lambda)$ sea localmente maximizada usando un procedimiento iterativo como el método Baum-Welch.

Para describir el procedimiento de reestimación de los parámetros HMM, primero se define el parámetro $\xi_t(i, j)$ como la probabilidad de estar en el estado S_i en un tiempo t, y en el estado S_j en un tiempo $t+1$ dada la secuencia de observaciones y el modelo; esto es

$$\xi_t(i, j) = P(q_t = S_i, q_{t+1} = S_j \mid O, \lambda) \qquad \text{Ec 5. 51}$$

Previamente se definió $\gamma_t(i)$ como la probabilidad de estar en el estado S_i en el tiempo t, dada la secuencia de observaciones y el modelo; entonces se puede relacionar $\gamma_t(i)$ con $\xi_t(i,j)$. Al Sumar sobre j, se obtiene que:

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \qquad \text{Ec. 5. 52}$$

La suma de $\gamma_t(i)$ y $\xi_t(i,j)$ sobre el tiempo t puede interpretarse como:

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{número esperado de transiciones desde } S_i \qquad \text{Ec. 5. 53}$$

$$\sum_{t=1}^{T-1} \xi_t(i,j) = \text{número esperado de transacciones desde } S_i \text{ hasta } S_j \qquad \text{Ec. 5. 54}$$

Usando las fórmulas anteriores, un conjunto razonable de fórmulas de reestimación de los parámetros de un HMM son:

$$\bar{\phi}_i = \text{número esperado en el estado } S_i \text{ en el tiempo (t=1)} = \gamma_1(i) \qquad \text{Ec. 5. 55}$$

$$\bar{a}_{ij} = \frac{\text{número esperado de transiciones del estado } S_i \text{ al estado } S_j}{\text{número esperado de transiciones del estado } S_i} = \frac{\sum_{t=1}^{T} \xi_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \qquad \text{Ec. 5. 56}$$

$$\bar{b}_j(k) = \frac{\text{número de veces esperado en el estado j y el símbolo observado } v_k}{\text{número de veces esperado en el estado j}} = \frac{\sum_{t=1}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \qquad \text{Ec. 5. 57}$$

Esta demostrado que el modelo inicial λ define un punto crítico de la función de similitud, por lo que $\overline{\lambda} = \lambda$, de otra forma el modelo $\overline{\lambda}$ es más parecido que el modelo λ en el sentido que $P(O \mid \overline{\lambda}) > P(O \mid \lambda)$.

Basado en el procedimiento anterior , si se usa iterativamente $\overline{\lambda}$ en vez de λ, podemos mejorar la probabilidad de que O sea observada desde el modelo. El resultado final se llama estimación de máxima similitud (ML) del HMM y proporciona un máximo local.

Para aplicar los resultados anteriores se necesita una escala para ambos problemas; se observa que $\alpha_t(i)$ y $\beta_t(i)$ consisten en numerosos términos pequeños. Para observaciones múltiples en el entrenamiento (problema 1), las fórmulas de reestimación se modifican al agregar todas las frecuencias individuales de ocurrencia para cada secuencia.

Un aspecto clave en el procedimiento de reestimación son las estimaciones iniciales de un HMM. No existe una respuesta única, pero por experiencia tanto una estimación aleatoria o uniforme de la ϕ y de la A son adecuadas. Sin embargo, para estos y para los parámetros B, un buen estimador inicial, es una variante del procedimiento de K-medias.

El procedimiento de estimación llamado de máxima similitud (ML), usado para la estimación del conjunto de parámetros de HMM, maximiza localmente la densidad de probabilidad del modelo para la señal a partir de los datos de entrenamiento. Otro algoritmo, el cual es referido como el algoritmo de segmentación de K-medias ha sido ampliamente utilizado para reconocimiento continuo y conectado[29][77]. Este algoritmo maximiza localmente la densidad conjunta del estado de secuencia y de la señal a partir de los datos de entrenamiento. La diferencia conceptual entre los dos algoritmos es que el algoritmo de Baum realiza la estimación con todas las posibles secuencias, mientras que el algoritmo de segmentación de K-medias considera solamente la mejor secuencia. Para reconocimiento de palabras aisladas, el algoritmo de segmentación de K-medias puede ser aplicado con un procedimiento de Viterbi[78]. Más que calcular los valores esperados de los eventos, los conteos se acumulan en cada estado basados en la secuencia de Viterbi Esto requiere mantener conteos para rastrear cada transición y cada símbolo de salida durante el entrenamiento.

Las ecuaciones de reestimación para el algoritmo de Viterbi son:

$$\hat{a}_{ij} = \frac{\text{número de transiciones del estado } S_i \text{ al estado } S_j}{\text{número esperado de transiciones del estado } S_i} \qquad \text{Ec. 5. 58}$$

$$\hat{b}_{ij}(k) = \frac{\text{número de transiciones del estado j y el símbolo } v_k}{\text{número esperado de veces en el estado j}} \qquad \text{Ec. 5. 59}$$

Además de la función de similitud (ML), otro criterio estima la mínima probabilidad de error[79]; sin embargo, este procedimiento es mucho más complejo con resultados cuestionables. Un método más formal es el de Máxima Información Mutua (MMI)[9] . Su objetivo es determinar los parámetros del modelo mediante la maximización de la probabilidad de generar un dato dada la secuencia correcta de palabras, como en el ML, pero al mismo tiempo, *minimizando la probabilidad de generar cualquier secuencia incorrecta de palabras*[80]. *Se han* obtenido resultados semejantes, pero el entrenamiento de MMI es más robusto cuando el modelo es incorrecto[81]. Un método diferente, el de Información del Mínimo Discriminante (MDI) se ha propuesto como generalización tanto de ML como de MMI[82].

Una desventaja con los HMM discretos es que para algunas aplicaciones las observaciones son continuas y se degradan en el proceso de cuantización. *Se podría mejorar* en el caso de vocabularios grandes con la aplicación de HMM con las densidades de observación continuas, teniendo en contraparte una mayor complejidad en los modelos.

Asumiendo que se tiene un vocabulario de V palabras por reconocer, cada palabra se *modela con un HMM. Para cada palabra (por uno o más locutores) donde cada repetición* constituye una secuencia de observaciones, y donde las observaciones son algunas representaciones apropiadas del espectro o de las características temporales de las palabras.

Para cada palabra v en el vocabulario, debemos construir un λ^v HMM; esto es, se deben estimar los parámetros del modelo que optimen la similitud del conjunto de observaciones de entrenamiento con la v-ésima palabra.

El cálculo de probabilidades se lleva a cabo mediante el algoritmo Viterbi. Para reconocimiento de palabras aisladas con un HMM más apropiado que el modelo ergódico, Los estados se relacionan con subpalabras de manera muy directa.

Así, el significado físico de los estados del modelo puede ser los fonemas, las sílabas o la palabra completa. Los modelos de Markov de primer orden son obviamente una limitación del modelo que en ocasiones es inapropiada para algunos sonidos. Se ha utilizado con una pequeña mejoría[83] un HMM discreto de segundo orden, extendiendo el algoritmo de Baum-Welch y las fórmulas de reestimación.

5.3.2.Método de clasificación

Entre los tipos de modelos ocultos de Markov, se utilizaron para este trabajo los discretos

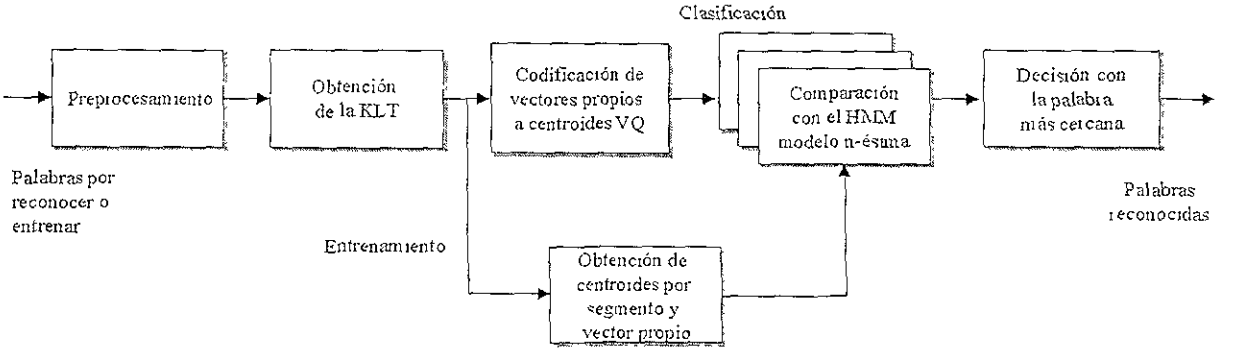


Figura 5 12. Diagrama de bloques del sistema de reconocimiento de palabras aisladas usando HMM.

de izquierda a derecha, usando las subpalabras como estados y generando centroides por estado para considerarlos como observaciones del modelo, y se aplicó el algoritmo de Viterbi como criterio para seleccionar el modelo de Markov más cercano El sistema utilizado se muestra en la figura 5.12. Para la fase de entrenamiento, se utilizaron 1,000 repeticiones de los dígitos para obtener los modelos de cada palabra. El sistema genera 16 centroides por estado y por vector propio, y se asumen valores no nulos para las probabilidades de estados iniciales distintos al primero.

En la clasificación, dado que cada palabra se compara con un modelo de Markov, sus subpalabras son codificadas a los centroides ya obtenidos en el entrenamiento. En esta etapa, ver figura 5.12, la decisión se toma al seleccionar la mayor probabilidad obtenida con cada modelo, las probabilidades resultan de aplicar el algoritmo de Viterbi a la palabra de prueba y los modelos de Markov

En la fase de clasificación se utilizaron 1,000 repeticiones de la base TI, los resultados para los parámetros KLT se expresan en la tabla 5.10. Si bien resultan muy bajos comparados con los expresados en las técnicas ya presentadas, coinciden con los resultados obtenidos anteriormente[14].

	0	1	2	3	4	5	6	7	8	9
0	88	3	2	1	2	0	0	0	0	4
1	1	90	2	0	1	0	0	0	0	6
2	3	1	94	0	2	0	0	0	0	0
3	1	0	3	89	0	4	3	0	0	0
4	1	2	2	0	90	3	0	0	2	0
5	2	0	0	0	4	91	0	0	0	3
6	0	0	0	3	0	3	88	4	0	2
7	3	0	1	3	0	0	3	86	2	2
8	0	0	0	0	4	0	3	0	93	0
9	2	8	0	1	0	0	0	1	0	88

Tabla 5.11. Matriz de confusión para palabras aisladas en inglés de la base de TI, usando HMM de 4 estados con parámetros KLT, precisión 89.7%.

	0	1	2	3	4	5	6	7	8	9
0	90	1	2	0	3	2	0	0	0	2
1	0	89	2	0	0	0	0	2	0	7
2	5	2	91	0	0	0	0	0	2	0
3	0	0	1	96	0	2	1	0	0	0
4	2	1	3	0	92	2	0	0	0	0
5	4	0	0	0	3	89	0	0	0	4
6	1	0	0	4	0	3	91	1	0	0
7	3	0	1	2	0	2	4	88	0	0
8	0	0	0	0	6	0	0	0	94	0
9	0	5	0	0	0	0	1	0	0	94

Tabla 5.12. Matriz de confusión para palabras aisladas en inglés de la base de TI, usando VQ de 4 segmentos, con *coeficientes* LPC, *precisión* 91.4%.

Para las otras técnicas de codificación, los resultados no mejoraron significativamente; así para los coeficientes LPC se obtuvo una precisión de 91.4%, ver tabla 5.12. Se observa que si bien los centroides pierden precisión, al obtener sólo probabilidades discretas de las observaciones, no se discrimina entre fonemas adecuadamente. Por lo anterior entre otras razones, los sistemas actuales de reconocimiento continuo construyen modelos de Markov por fonema o subpalabra y no por toda la palabra.

En el caso de la base de datos en español, el ruido ambiente fue crucial para que el sistema se comportara de manera mucho más imprecisa, teniendo precisiones abajo de 80%, de lo que se concluye que no es útil esta técnica con el uso de subpalabras como unidades de reconocimiento.

Capítulo 6

RECONOCIMIENTO DE PALABRAS CONECTADAS

El problema de reconocer palabras de manera fluida en un medio natural fue el primer objetivo de los investigadores en esta área[6]. Como ya se ha mencionado, en virtud de la dificultad de este problema, las investigaciones se concentraron en el reconocimiento de palabras aisladas, después en conectadas y, finalmente, en palabras continuas. En el reconocimiento de palabras aisladas se obtuvieron precisiones altas en medios naturales en la década de los 80, ahora el objetivo en la última década ha sido mejorar estas precisiones y hacerlas extensivas a palabras conectadas y continuas, para esto se han utilizado las técnicas DTW y HMM.

Un problema que aparece en el reconocimiento de palabras conectadas y continuas, y no el de aisladas, es la coarticulación. Es decir, la modificación de fonemas por la presencia de sonidos vecinos a éste.

Para el reconocimiento de palabras conectadas se ha usado una gran cantidad de unidades, menores a una palabra, pero que a su vez no sean tan numerosas, las más utilizadas son de tipo lingüístico como fonemas, difonemas, sílabas, demisílabas y disílabas.

Las principales tareas para un exitoso reconocimiento de palabras conectadas son: identificar el número de unidades en una frase y dónde se localizan sus límites, la clasificación eficiente de cadenas de palabras, y el realizar un entrenamiento adecuado.

Las principales técnicas para las dos primeras tareas son: dos niveles de empalme DP por Sakoe[19], el edificio de niveles por Myers y Rabiner[20], y un solo paso DP por Vintsyuk[23], Bridle y Brown[24]. Estas técnicas se han utilizado como métodos de clasificación para DTW y HMM.

Para mejorar la clasificación DTW se han utilizado muchas técnicas de entrenamiento[9], entre las que destacan: el entrenamiento robusto[25], el entrenamiento embebido[27], éste término del ingles *embedded*, y el entrenamiento por k-medias[29].

En los dos siguientes subtemas se describirán las técnicas de clasificación y entrenamiento mencionadas. En el tercer y último subtema se describirá el método propuesto que consiste en la técnica de un solo paso con unidades acústicas caracterizadas por los parámetros KLT, se comparará su desempeño con coeficientes LPC como unidades; como entrenamiento se utilizaron palabras aisladas de la misma base de datos de cadenas de palabras, y se comparará su desempeño cuando el entrenamiento se obtiene de palabras que se aíslan de cadenas de palabras.

6.1 Clasificación de palabras conectadas

En esta sección, se tratarán las primeras dos metas presentadas en la introducción: reconocer el número de unidades en una frase y dónde se encuentran las fronteras, y comparar eficientemente cadenas de palabras.

En el reconocimiento aislado de voz, el principio y fin de la palabra son razonablemente fáciles de detectar. Para palabras conectadas, la segmentación es difícil ya que las fronteras entre palabras no son fáciles de obtener debido a la coarticulación natural de la voz. Así que las técnicas de palabras aisladas que requieren presegmentación no pueden ser aplicadas al reconocimiento de palabras conectadas. La mejor estrategia de reconocimiento de palabras conectadas es optimar simultáneamente la segmentación y el reconocimiento, en vez de hacerlo sucesivamente[84]. En las técnicas DTW se optiman simultáneamente la segmentación y el reconocimiento.

En principio, cada posible segmentación es considerada para todas las posibles secuencias de palabras. Si aplicamos esta búsqueda exhaustivamente, entonces los cálculos son excesivos. La idea básica de las técnicas DTW es cómo reducir las rutas de búsqueda. Se revisarán las tres técnicas más utilizadas que ya se mencionaron. De estas tres aproximaciones, la de un solo paso[22] no requiere de varios niveles de optimación, y su simplicidad permite una realización eficiente y mejor comportamiento en tiempo real. Las técnicas DTW para reconocimiento de palabras conectadas se pueden extender a HMM, para este último método se revisará la técnica del edificio de niveles[65]. Las técnicas DTW para el reconocimiento de palabras conectadas se describirá a continuación.

6.1.1 Técnicas DTW

Supóngase que tenemos una cadena de patrones $T = T(m)$ de M tramas y de un número desconocido de palabras. La cadena T se considera como N secuencias de L patrones de referencia $R_{q(1)}(n), R_{q(2)}(n), \ldots, R_{q(L)}(n)$, donde cada patrón está dentro de un conjunto de V patrones ; $R_q, q = 1,2,\ldots,V$; $n = 1,2,\ldots,N_q$; y $L_{MIN} \leq L \leq L_{MAX}$. Se define un super patrón de referencia $R'(n)$ como la concatenación de L patrones de referencia:

$$R^s(n) = R_{q(1)}(n) \oplus R_{q(2)}(n) \oplus \ldots \oplus R_{q(L)}(n) \qquad \text{Ec 6. 1}$$

El patrón de super referencia $R^s(n)$ puede ser ahora expresado como:

$$R^s(n) = \begin{cases} R_{q(1)}(n-\phi(0)), & 1+\phi(0) \le n \le \phi(1) \\ R_{q(2)}(n-\phi(1)), & 1+\phi(1) \le n \le \phi(2) \\ \vdots & \\ R_{q(L)}(n-\phi(L-1)), & 1+\phi(L-1) \le n \le \phi(L) \end{cases} \qquad \text{Ec 6. 2}$$

donde la longitud de la función $\phi(1)$ esta definida como

$$\phi(1) = \sum_{k=1}^{1} N_{q(k)} \quad \text{and} \quad \phi(0) \qquad \text{Ec 6. 3}$$

donde N_q tiene una longitud q-ésimo de patrón de referencia.

El registro o acoplamiento del patrón de prueba T con un patrón de super referencia $R^s(n)$ puede ser propuesto como un problema de DTW, esto se ilustra en la figura 6.1.

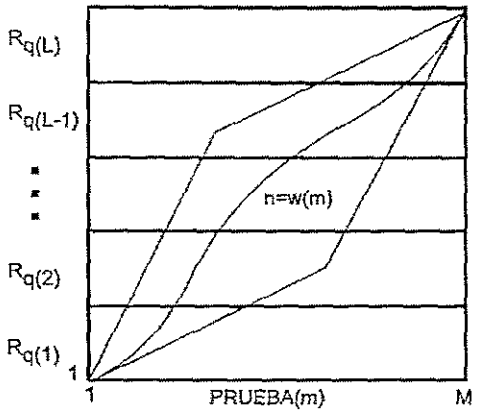


Figura 6.1. Acoplamiento DTW para palabras conectadas.

La ruta óptima de acoplamiento $n = w(m)$ es un mapeo entre la trama de prueba m y la superficie de referencia n. Asumiendo L y $q(1) = 1,2,\ldots,L$, son conocidos, $R^s(n)$ se puede considerar como un patrón de referencia simple, y el alineamiento entre T y $R^s(n)$ puede

calcularse como cualquier DTW. La distancia local entre la trama m-ésima del patrón de prueba y la n-ésima de la super referencia se denota mediante $d(m,n)$. La distancia global D En el reconocimiento de palabras aisladas es:

$$D = \min_{w(m)} \left[\sum_{m=1}^{M} d(m, w(m)) \right] \qquad \text{Ec 6. 4}$$

la cual puede estar sujeta a los limites de inicio y fin de punto $w(1)=1$ y $w(M)=\phi(L)$, o bien a algunas pendientes globales. El elemento común sigue siendo la continuidad, que en el caso más limitante es cuando no se evita ninguno de los puntos:

$$m(k)-m(k-1)\leq 1 \quad \text{y} \quad n(k)-n(k-1)\leq 1 \qquad \text{Ec 6. 5}$$

Las trayectorias locales consideradas para palabras aisladas, ver figura 5.2, se vuelven a aplicar a palabras conectadas, entre las que destacan las de Itakura, inciso (h) en la figura 6.2, de Sakoe y Chiba mostradas en los incisos (a) a (d) de la misma figura, y Myers y Rabiner, en los incisos (e) a (g)[71].

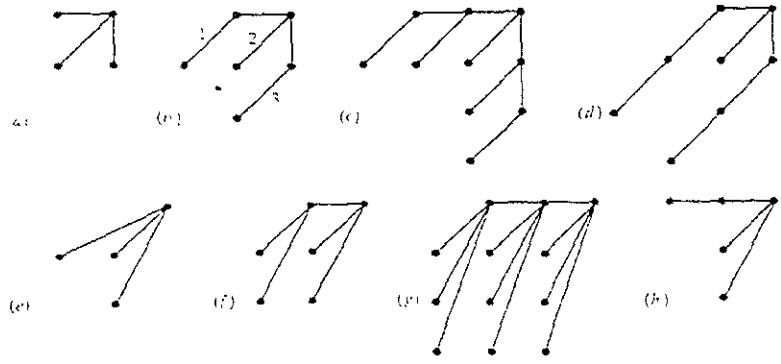


Figura 6.2. Trayectorias locales aplicadas a reconocimiento de palabras conectadas, Itakura (h), Sakoe (a)-(d) y Myers (e)-(g)[71]

Para llegar a una comparación irrestricta entre T y $R^{\wedge}$, primeramente se fija a L y $q(l)$ para toda l, obteniendo distancias resultantes $D_{q(1)q(2)\ \ q(L)}(M)$, así por cada L, encontramos una combinación óptima de patrones parciales de referencia $q(1)q(2),\ldots,q(L)$. Finalmente, la solución D^{*} es la minimización sobre todos los posibles valores de L, donde $L_{MIN} \leq L \leq L_{MAX}$

$$D^* = \min_{L}\left[\min_{q(1)q(2)\ldots q(L)}\left[D_{q(1)q(2)\ldots q(L)}(M)\right]\right] \qquad \text{Ec 6. 6}$$

Se observa que este esquema de reconocimiento no necesita segmentar la cadena de prueba. Esta labor directa requiere de cálculos excesivos, esto es, el número de secuencias para $L_{MAX} = V$ y $L_{MIN} = 1$ es $\sum_{L=1}^{V} P_L^{V} = \sum_{L=1}^{V} V^L$ donde P_L^V denota la permutación de L objetos del conjunto de V objetos.

Por ejemplo, para $L = 6$ y $V = 12$ (un vocabulario de 12 palabras y con más de 6 palabras encadenadas), se tienen que $12^6 + 12^5 + \ldots + 12$ secuencias deben de ser probadas. Los métodos que se presentarán reducen la carga de cómputo resolviendo la minimización por etapas o niveles, por ejemplo por palabras, en donde se retiene suficiente información. Así, la mejor cadena se retiene como una de las candidatas y la mayoría de cadenas con distorsiones altas son descartadas.

Los pasos en el método DTW[20] para reconocimiento aislado de palabras son los siguientes:

Inicialización $D_{\min}(0,0) = 0, D_{\min}(0,n) = \infty, n \neq 0$.

Iteración, para $m = 1,2,\ldots,M,$, calcular

$$D_{\min}(m,n) = d(m,n) + D_{\min}(m,\hat{n}) \qquad \text{Ec 6. 7}$$

Solución final

$$D^* = D_{min}(M,N) \qquad \text{Ec 6. 8}$$

donde: $D_{\min}(m,n)$ es la distancia acumulada a lo largo de la mejor forma de ruta a partir del punto inicial hasta las tramas m^{th} y n^{th} del patrón de prueba y de la super referencia, respectivamente; $D_{\min}(m,\hat{n})$ es la mínima distancia sobre todas las rutas permitidas antes de los puntos (m,n).

Ahora el problema en el reconocimiento de palabras conectadas, es cómo introducir en este procedimiento un número variable de niveles (palabras) que retienen a las mejores cadenas.

Algoritmo de dos niveles

La minimización del problema se divide en dos pasos o niveles: uno para el nivel de palabras, y otro para todas las palabras conectadas (nivel de frase).

Dado un patrón de prueba T de M vectores trama, $T=\bar{t}_1,\bar{t}_2,\ldots,\bar{t}_M$ y un patrón de superreferencia $R^{\wedge}$. Se dividen en L patrones parciales:

$$T=T_1\oplus T_2\oplus\ldots\oplus T_L$$

$$R^{\wedge}=R_{q(1)}\oplus R_{q(2)}\oplus\ldots\oplus R_{q(L)} \qquad \text{Ec 6. 9}$$

Asumiendo que T tiene $(L-1)$ fronteras $b_1,b_2,\ldots,b_{L-1}$. El patrón de prueba puede ser expresado como:

$$T=T_1(b_0+1,b_1)\oplus T_2(b_1+1,b_2)\oplus\ldots\oplus T_L(b_{L-1}+1,b_L) \qquad \text{Ec 6. 10}$$

El patrón parcial T_i puede ser expresado en términos de su trama final b_i y su número de tramas m_i, entonces $T_i(b_{i-1}+1,b_i)=T(b_i,m_i)$.

La distancia mínima está dada por:

$$D^{*}=\min_{L,b_i}\left\{\sum_{i=1}^{L}\min_{q(i)}\left[D(T(b_i,m_i),R_{q(i)})\right]\right\} \qquad \text{Ec 6. 11}$$

La minimización interna es el acoplamiento en el nivel de palabra. Para esta combinación debemos calcular un mínimo para cada valor de L y de b_i, lo que se denota por:

$$\min_{q(i)}\left[D(T(b_i,m_i),R_{q(i)}\right]=\min_{q}D(b_i,m_i,q)=\hat{D}(b_i,m_i) \qquad \text{Ec 6. 12}$$

Aquí se encuentra el conjunto óptimo de parámetros $\hat{q}$ para cada patrón parcial. Esta minimización es equivalente al reconocimiento de palabras aisladas. Mucho del cálculo se realiza en este nivel, así que es conveniente evitar duplicar cálculos de las mismas distancias.

El alineamiento en el nivel de frase resuelve la minimización

$$\min_{L}\left\{\min_{b_i}\left|\sum_{i=1}^{L}\hat{D}(b_i,m_i)\right|\right\} \qquad \text{Ec 6. 13}$$

de donde se obtienen los parámetros óptimos $\hat{b}_i$ y $\hat{L}$, hacia atrás se obtiene el conjunto de parámetros $\hat{q}$.

Algoritmo edificio de niveles

El algoritmo DTW se desarrolla en niveles, esto es, una referencia a la vez, como se observa en al figura 6.3.

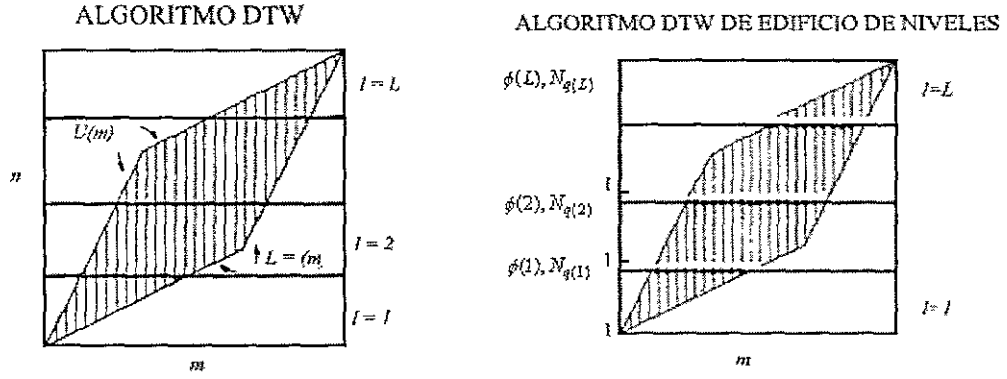


Figura 6.3. Algoritmo edificio de niveles y su comparación con el DTW de palabras aisladas,

Para el algoritmo DTW en palabras aisladas, el cálculo se hace en líneas verticales (así un valor de n para cada m) las cuales generalmente incluyen tramas de dos o más referencias. Para el edificio de niveles, el cálculo se hace mediante una secuencia de líneas verticales dentro de cada referencia. Siguiendo todas las líneas verticales para la referencia l èsima, el conjunto de distancias acumuladas a lo largo de la horizontal $n=\phi(l)$ es almacenado (como se muestra con los puntos sólidos), y usado como distancia inicial para el siguiente nivel. De esta forma, sigue el cálculo para otros niveles.

El algoritmo edificio de niveles se puede extender cuando patrones de prueba son variables, cuando los fines de palabra se desconocen, y el algoritmo agrega un procedimiento hacia atrás para obtener los fines de palabra. Un resumen del algoritmo anterior se presenta a continuación:

$D_l^q(m)$, la distancia acumulada en el nivel l del patrón de referencia q, a la trama m;

$F_l^q(i)$, un apuntador de regreso indicando donde comenzó la ruta al principio de nivel, en el nivel l, del patrón de referencia q, a la trama i.

Una vez que $D_l^q(m)$ está calculada para toda q, se lleva a cabo una minimización sobre q para obtener al mejor elemento, $W_l^B(m)$,en cada trama m, así como también la distancia acumulada $D_l^B(m)$, y el apuntador $F_l^B(m)$ que corresponde al mejor elemento, para el nivel l,como sigue:

$$\begin{aligned} D_l^B(m) &= \min_{1\le q\le V} D_l^q(m), \qquad 1 \le m \le M, \\ W_l^B(m) &= \arg\min_{1\le q\le V} D_l^q(m), \quad 1 \le m \le M, \\ F_l^B(m) &= F_l^{W_l^B(m)}, \qquad 1 \le m \le M, \end{aligned} \qquad \text{Ec 6. 14}$$

esto concluye el nivel 1. El procedimiento es iterativo en los niveles siguientes tomando las rutas de DTW a partir de cualquier trama m, donde $1 \le m \le M$, dentro de la región final del nivel que le precede.

La cadena óptima está determinada por el nivel l, para el cual la distancia acumulada al final de la cadena es mínima, entonces se encuentra la secuencia de la palabras que corresponde a la cadena óptima usando $W_l^B(m)$ y mediante el rastreo de hacia atrás hasta principio de la cadena.

Los puntos de segmentación óptima de T en palabras individuales se obtienen usando $F_l^B(m)$ y rastreando hacia atrás. Si el número de palabras en la cadena es conocida, la cadena óptima de esa longitud se obtiene mediante la observación de las salidas en el nivel que corresponda al número conocido de palabras.

En términos del algoritmo de Sakoe, intercambiamos el orden de minimización

$$D^* = \min_{L,b_i}\left\{\sum_{i=1}^{L}\min_{q(i)}\left[D(T(b_i,m_i),R_{q(i)})\right]\right\} \qquad \text{Ec 6. 15}$$

por

$$D^* = \min_{L}\left\{\min_{q(i)}\left[\sum_{i=1}^{L}\min_{b_i} D(T(b_i,m_i),R_{q(i)})\right]\right\} \qquad \text{Ec 6. 16}$$

calculando para toda b_i, y para un valor fijo de $q(i)$, la distancia $D(T(b_i,m_i),R_{q(i)})$ en un solo aclopamiento. Esto reduce los cálculos en un orden de magnitud.

Algoritmo de un solo paso

Este algoritmo realiza una parametrización de las rutas con un solo índice y trata el criterio de optimización directamente como si fuera una función de esta ruta de alineamiento. Este método es mucho más sencillo y eficiente que los dos antes mencionados.

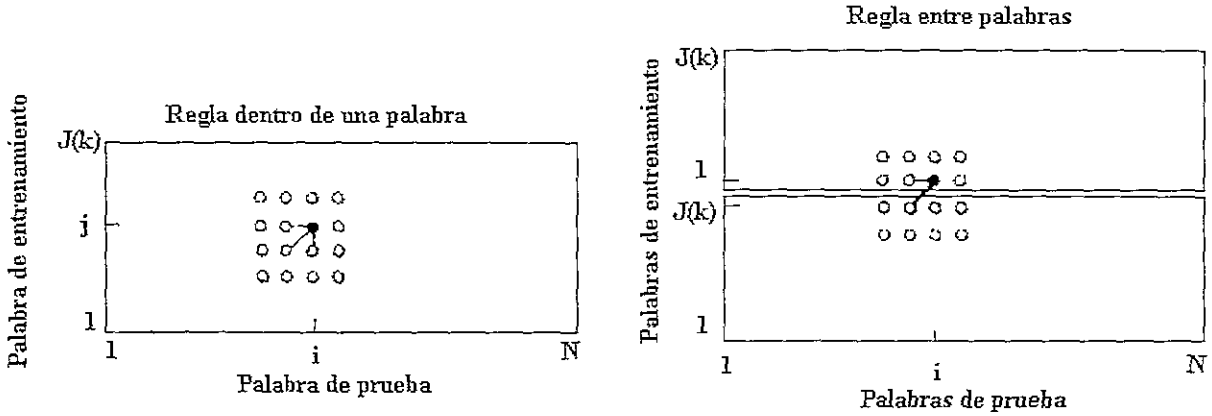


Figura 6.4. Tipos de trayectorias locales dentro de palabras y en fronteras de palabras[22],

Estas técnicas utilizan dos tipos de trayectorias locales: dentro de las palabras y en sus fronteras. Una rejilla en un punto de la ruta se denota como (m,n,q), donde m es una trama del patrón de prueba, n es una trama del patrón de referencia q. Los límites usados por Ney[22] son:

1. Dentro de palabras,si $w(q)=(m,n,q), n>1$, entonces

$$w(q-1) \in \{(m-1,n,q),(m-1,n-1,q),(m,n-1,q)\}. \qquad \text{Ec 6. 17}$$

2. En las fronteras de palabras, donde $n=1$, si $w(q)=(m,n,q)$ entonces

$$w(q-1) \in \{(m-1,n,1);(m-1,L(q^*),q^*): q^* = 1,\ldots,V\}. \qquad \text{Ec 6. 18}$$

donde $L(q^*)$ es la longitud de la trama de la palabra q^* que precede a la palabra q. Estos limites se muestran en la figura 6.4.

Esta técnica agrega al algoritmo DTW de palabras aisladas la regla (2) para las fronteras de palabras y selecciona la distancia mínima acumulada, en cualquier punto a lo largo de todas las palabras del vocabulario.

	Algoritmo de dos niveles	Algoritmo de edificio de niveles	Algoritmo de edificio de niveles modificado	Algoritmo de un solo paso
Número de alineamientos	KN	KM	KM	K
Tamaño de los alineamientos	$\bar{J}(2R+1)$	$\bar{J}(N/3)$	$\bar{J}(2R+1)$	$\bar{J}N$
Total de cálculos	$\bar{J}KN(2R+1)$	$\bar{J}KM(N/3)$	$\bar{J}KM(2R+1)$	$\bar{J}KN$
Almacenamiento	$2N(2R+1)$	$3NM$	$3NM$	$2(\bar{J}+K+N)$
Número de alineamientos	*3,600*	*120*	*120*	*10*
Tamaño de los alineamientos	*875*	*4,200*	*875*	*12,600*
Total de cálculos	*3'150,000*	*504,000*	*105,000*	*126,000*
Almacenamiento	*18,000*	*12,960*	*12,960*	*1,420*

Tabla 6 1. Comparación de los métodos DTW para palabras conectadas[22].

El algoritmo es entonces el siguiente:

1. Se calcula

$$D(1,n,q) = \sum_{i=1}^{n} d(1,i,q). \qquad \text{Ec 6. 19}$$

2. a) Para $m = 2,\ldots,M$, realizar pasos 2b-2e.

b) Para $q = 1,\ldots,V$ realizar pasos 2c-2e.

c) Se calcula

$$D(m,1,q) = d(m,1,q) + \min\{D(m-1,1,q); D(m-1,L(q^*),q^*) : q^* = 1,2,\ldots,V \qquad \text{Ec 6. 20}$$

d) Para $n = 2,\ldots,L(q)$, realizar paso 2e.

e) Se calcula

$$D(m,n,q) = d(m,n,q) + \min\{D(m-1,n,q), D(m-1,n,q), D(m,n-1,q)\}. \qquad \text{Ec 6. 21}$$

3. Obtener hacia atrás la mejor ruta usando el arreglo $D(m,n,q)$ de las distancias acumuladas.

En la tabla 6.1[22] se muestra la comparación en general y para un caso numérico entre los tres algoritmos descritos y un tipo mejorado del edificio de niveles[21], este último resulta del del algoritmo del nivel restringiendo la región de búsqueda para la mejor ruta, donde: K es el número de palabras, y se obtiene para $K = 10$; $\bar{J}$ es la longitud promedio de la palabra, se utilizó $\bar{J} = 35$; M es número máximo de palabras en la cadena de entrada, se usó $M = 12$; N es la longitud de la cadena de entrada, se usó $N = 360$; y R es el rango para el tiempo de alineamiento, se usó $R = 12$.

El algoritmo de un solo paso requiere solamente 1/25 de cálculos del algoritmo de dos niveles, 1/4 del algoritmo de edificio de dos niveles , y cálculos semejantes del algoritmo modificado del edificio de dos niveles. Respecto al almacenamiento, el algoritmo de un solo paso ofrece un factor de reducción en un factor de 9 o más comparado con los otros tres algoritmos.

6.1.2 Técnica HMM

Fundamentalmente el proceso es casi el mismo que el utilizado para el algoritmo edificio de niveles. Denotemos a la cadena de prueba como $O = O_1, O_2 \ldots O_M$ y al conjunto de modelos de Markov de V palabras como M_q, $1 \le q \le V$. Para encontrar la secuencia optima de HMM's

que se ajuste a O, aplicamos a cada HMM M_q un algoritmo de Viterbi con O, empezando en la trama 1, nivel 1 y reteniendo para cada posible trama m, lo siguiente[85]:

1. $P_l^q(m)$, esto es la probabilidad acumulada a la trama m, en el nivel l para el modelo de referencia M_q, a lo largo de la mejor ruta.

2. $F_l^B(m)$, un apuntador hacia atrás indicando donde principia la ruta en el comienzo del nivel.

La medida local para calcular $P_l^q(m)$ utilizada por Myers y Rabiner es:

$$P_l^v(O) = b_l^v(O_t)\left[p_l^v(\varepsilon_t)\right]^{\gamma_\varepsilon} K_1 \qquad \text{Ec 6. 22}$$

donde γ_ε es un coeficiente de energía escalar, Rabiner[85] utiliza el valor 0.375, y K_1 es una constante de normalización que depende de γ_ε, de forma tal que la última ecuación es una probabilidad. Las probabilidades de transición ingresan al cálculo de $P_l^q(m)$ vía la optimización por el proceso de Viterbi.

Como en el caso de palabras aisladas, al final de cada nivel, l, una maximización sobre q se lleva a cabo para obtener al mejor modelo en cada trama m como sigue:

$$\begin{aligned} P_l^B(m) &= \max_{1\le q\le V} P_l^q(m), & 1 \le m \le M, \\ W_l^B(m) &= \arg\max_{1\le q\le V} P_l^q(m), & 1 \le m \le M, \\ F_l^B(m) &= F_l^{W_l^B(m)}, & 1 \le m \le M, \end{aligned} \qquad \text{Ec 6. 23}$$

La mejor cadena de l palabras $(1 \le l \le L)$ con probabilidad $P_l^B(M)$ se obtiene procesando hacia atrás con el apuntador $F_l^B(m)$ que señala las palabras en la cadena. La mejor cadena es el máximo de $P_l^B(M)$ sobre todos los posibles niveles, l.

6.2 Métodos de entrenamiento

Los primeros reconocedores de palabras conectadas utilizaban palabras aisladas en el entrenamiento[6][86]. Esta es la forma más sencilla de derivar los patrones de palabras completas usados en el reconocimiento de palabras conectadas.

Existen al menos dos problemas inherentes al comparar secuencias de palabras conectadas con las concatenadas, estas últimas se forman con palabras aisladas. Primero, las palabras aisladas tienden a ser de mayor duración que las mismas inmersas en cadenas conectadas. Segundo, las palabras conectadas en cadenas coarticulan en las fronteras, por lo que producen un comportamiento espectral significativamente diferente, que por cada palabra individual.

Los algoritmos DTW pueden compensar diferencias de duraciones hasta cierto punto, así el entrenamiento por palabras aisladas trabaja razonablemente bien en cadenas articuladas de palabras lentas, y aún en cadenas de palabras con velocidad moderada (100 palabras por minuto, ppm) donde la coarticulación sea pequeña. Sin embargo, a partir de las tasas de voz de alrededor de 150 ppm las palabras coarticuladas son sustancialmente más pequeñas que las aisladas, y en consecuencia los patrones de entrenamiento son inadecuados. También, debe considerarse que la compresión de palabras en frases es no lineal, por ejemplo, para las vocales y fricativas sus regiones estables son reducidas y/o recortadas, mientras que sus regiones de transición son poco afectadas.

Para remediar los problemas asociados con el uso de patrones de entrenamiento de palabras aisladas, tres técnicas principales han sido utilizadas para clasificación DTW: entrenamiento robusto [25][26], palabras inmersas[27][28], y segmentación k-medias[29][30]. La primera utiliza un entrenamiento modificado de palabras aisladas, la segunda combina patrones de referencia de palabras aisladas con referencias de palabras extraídas de cadenas de palabras conectadas. La última utiliza palabras extraídas de cadenas de palabras , y desarrolla modelos de palabras estrictamente tomados de éstas. Se describirán estas técnicas, enfatizando los resultados obtenidos al aplicarlas.

6.2.1 Entrenamiento robusto

En el procedimiento robusto cada palabra se representa por una referencia. Ésta se obtiene al promediar un par de repeticiones semejantes. Se aplica una DTW y se promedian valores, incluida la duración. Las repeticiones contaminadas con algún tipo de ruido son eliminadas[25].

Primeramente, el entrenamiento robusto se utilizó en sistemas de reconocimiento de palabras aisladas. Es más confiable y robusto que el entrenamiento multirreferencia (en éste, el hablante, repite cada palabra del vocabulario varias veces, y se crea una variable de referencia por cada palabra), y tiene ventajas sobre el promedio (aquí, las repeticiones de cada palabra son promediadas y se obtiene una referencia por palabra), como sobre el cuantizador vectorial (aquí, las repeticiones son cuantizadas en un centroide por palabra).

Este procedimiento fue comparado con una sola o doble referencia por palabra en un sistema de reconocimiento de palabras aisladas, dando un error menor en 1% a 3% que con un entrenamiento multirreferencia[25]. Este procedimiento fue comparado también con una doble variación por palabra, usando comienzos variables y regiones terminales, para seis hablantes, cada uno habló 80 cadenas aleatoriamente, en un modo dependiente del hablante[26]. Los resultados se muestran en la tabla 6.2 con mejoras no concluyentes.

Precisión en el reconocimiento		
Cadenas de longitud	Dos referencias por palabra	Entrenamiento robusto
desconocida	*94 79%*	*95 21%*
conocida	*96 67%*	*96 25%*

Tabla 6 2. Comparación de precisiones en palabras conectadas, usando dos referencias fijas y el entrenamiento robusto[26].

6.2.2 Técnica embebida

La idea básica es mejorar el conjunto de referencias combinando patrones de referencia de palabras aisladas con patrones de palabras conectadas. Se debe contender con los tres

problemas conectados siguientes: (1) En secuencias de palabras conectadas, frecuentemente es difícil, casi imposible, asignar fronteras a las palabras de la cadena. Al incrementar la velocidad de la voz, este problema se agrava. (2) Existen palabras embebidas en cadenas con fuerte o débil coarticulación, cuyas estructuras espectrales son modificadas fuertemente. (3) Con la velocidad de voz del hablante se varía fuertemente la coarticulación de las palabras.

Está técnica de entrenamiento consiste en obtener un patrón de referencia por palabra. Para esto el sistema almacena cadenas de palabras conocidas y algunas palabras aisladas provenientes de cadenas. El propósito es que el último grupo se vaya ampliando a partir de cadenas de nuevas palabras.

Cuando una nueva cadena entra al sistema de entrenamiento, un algoritmo DTW la separa por palabras, usando las cadenas de referencia ya mencionadas. Las palabras de la nueva cadena son entonces comparadas con las palabras de referencia usando otro algoritmo DTW. De cada conjunto por palabra se escogen las dos repeticiones que tienen la mínima distancia entre ellas, y éstas se promedian en sus valores usando DTW, el promedio constituye la referencia para esa palabra[27].

Precisión en el reconocimiento				
Cadenas	Entrenamiento embebido y con palabras aisladas		Entrenamiento robusto	
longitud conocida	Cuidadosamente articuladas	Tasa normal	Cuidadosamente articuladas	Tasa normal
	93.13%	*85.63%*	*98.13%*	*98.13%*

Tabla 6.3. Comparación de precisiones en palabras conectadas, usando entrenamientos embebido y robusto[27].

Algunos resultados obtenidos en 1982[27] que se presentan en la tabla 6.3, muestran que las mejores precisiones se obtienen con las combinaciones de variaciones de palabras aisladas y embebidas. Los autores utilizaron una combinación de coarticulaciones normales y limitadas de palabras, en cadenas de 2 a 5 dígitos, cuatro hablantees produjeron las cadenas de entrenamiento.

Se probó el procedimiento para 16,17 y 18 hablantes, en modos tanto dependiente como independiente, pero solo se muestran en la tabla 6.3 los resultados de los 4 mejores. En un artículo subsecuente[28], los autores obtuvieron para 19 hablantes en modo dependiente un error en cadenas de longitud conocida de 5%, y de 7% para cadenas de longitud desconocida, de dos a cinco dígitos en las cadenas en ambos casos. Observaron que la inclusión de entrenamiento de dígitos embebidos mejora el reconocimiento para cadenas con velocidades y coarticulaciones normales, pero a su vez lo disminuye en cadenas de longitud desconocida. También obtuvieron que para el reconocimiento independiente del hablante, para la mayoría de los hablantes hubo pequeñas cadenas de tasa de error, 2.5% o menos, y para algunos hablantes la tasa de error fue de 50%.

6.2.3 Segmentación k-medias

El procedimiento de entrenamiento consiste en los siguientes pasos:

Se almacena un conjunto inicial de palabras aisladas, no importando si pertenecen o no a cadenas. El primer paso en el entrenamiento es segmentar las cadenas de entrenamiento en palabras aisladas utilizando cualquier algoritmo como el de niveles de construcción. El objetivo es tener una cantidad múltiple de repeticiones por palabra.

El segundo paso, con las repeticiones de cada palabra de entrenamiento, se realiza una cuantización vectorial si se están usando métodos DTW en la clasificación. Si el método de clasificación es HMM, entonces se aplica un método de reestimación a las repeticiones.

El último paso en este ciclo es la prueba para la convergencia. Si la diferencia entre el conjunto último de palabras de referencia y las palabras originales es pequeña, entonces el procedimiento termina. En caso contrario, las últimas palabras pasan a ser parte de las palabras originales de referencia y se vuelven a aplicar los dos últimos pasos, ya sea con nuevas palabras o con el conjunto que no había convergido.

Una aproximación más pragmática es utilizar un conjunto independiente de cadenas de palabras conectadas y almacenarlas en un archivo de prueba, para evaluar el comportamiento de reconocimiento del conjunto de referencia de palabras actualizadas; si la aproximación del

reconocimiento incrementa, el ciclo de entrenamiento es reiterativo; de otro modo, se asume que hay convergencia y el entrenamiento cíclico termina.

Precisión en el reconocimiento			
Cadenas de longitud	Entrenamiento embebido	Segmentación K-medias	
		VQ	HMM
desconocida	*81.3 %*	*98.7 %*	*97.8 %*
conocida	*84.7 %*	*99.0 %*	*99.4 %*

Tabla 6.4. Comparación de precisiones en palabras conectadas, usando entrenamientos embebidos y de segmentación k-medias[29].

El procedimiento de entrenamiento forma un ciclo cerrado, donde cada iteración mejora el objetivo. Un sistema con este procedimiento de entrenamiento fue comparado con una técnica embebida para el reconocimiento de palabras conectadas en 1986[29]. Para 4 hablantes en un modo dependiente y cadenas de longitud conocida y desconocida, se usaron 525 cadenas para entrenamiento y 525 para prueba, con los resultados mostrados en la tabla 6.4.

Precisión en el reconocimiento				
Cadena	Entrenamiento embebido y con palabras aisladas		Segmentación k-medias, HMM	
	dependiente	independiente	dependiente	independiente
longitud desconocida	*97.0 %*	*81.2 %*	*98.17 %*	*94.00 %*
longitud conocida	*98.4 %*	*85.5 %*	*99.19 %*	*96.60 %*

Tabla 6.5. Comparación de precisiones en palabras conectadas, usando un entrenamiento embebido y con palabras aisladas y segmentación k-medias con HMM[30].

De la tabla 6.4 puede observarse una significativa mejora en la precisión. Un año después, los mismos autores[30] obtuvieron los resultados mostrados en la tabla 6.5 para un HMM de 50 hablantes en un modo tanto dependiente como independiente, con este procedimiento de entrenamiento.

Otras técnicas han sido utilizadas para métodos de clasificación HMM, entre las que destaca la de algoritmos de agrupamiento[31], y un conjunto de modificaciones empleadas para incrementar la flexibilidad de este algoritmo. Algunas de estas últimas son: el uso de reglas heurísticas de la K-vecinos más cercana (KNN) para el reconocimiento independiente del hablante[26], técnicas que utilizan un libro de código LVQ[87] o información parcial de VQ[88], técnicas adaptadas para entrenamiento que incluyen estimación de la máxima información mutua[89][90], y un entrenamiento en bloques que adapta las nuevas palabras de referencia a otras ya almacenadas[91].

Nuevas técnicas de entrenamiento siguen siendo experimentadas. Para el propósito de esta investigación se utilizaron solamente las dos primeras, combinándolas con palabras aisladas.

6.3 Pruebas de reconocimiento

La base utilizada fue la de Texas Instruments para dígitos conectados en inglés, que contiene dos repeticiones de cada dígito aislado por hablante, además de 16 hablantes y 10 repeticiones por cada cadena de dígitos de éste número de hablantes, con cadenas de 2 a 7 dígitos, a excepción de 6 dígitos de los que no se tienen los datos.

longitud de la cadena en dígitos	Número de repeticiones	errores							
		0	1	2	3	4	5	6	7
2	156	124	27	5	0	0	0	0	0
3	148	105	34	7	2	0	0	0	0
4	153	99	35	11	5	3	0	0	0
5	160	92	38	13	7	6	4	0	0
7	153	61	52	18	12	7	2	1	0

Tabla 6.6. Matriz de confusión para palabras conectadas en inglés de la base de TI, método DTW de un solo paso, con parámetros KLT, entrenamiento con palabras aisladas.

Las pruebas se realizaron para los parámetros reducidos de la KLT y los coeficientes LPC, usando en ambos el método de un solo paso, en un modo dependiente del hablante. La variación realizada al algoritmo fue la de incorporar una vecindad de posibles punto de inicio y fin de palabras, que son utilizados en el algoritmo DTW.

longitud de la cadena en dígitos	Número de repeticiones	errores							
		0	1	2	3	4	5	6	7
2	156	130	22	4	0	0	0	0	0
3	148	101	39	5	3	0	0	0	0
4	153	103	37	7	5	1	0	0	0
5	160	95	40	15	4	4	2	0	0
7	153	66	49	15	12	6	3	2	0

Tabla 6.7. Matriz de confusión para palabras conectadas en inglés de la base de TI, método DTW de un solo paso, con coeficientes LPC, entrenamiento con palabras aisladas.

Para la pruebas se consideraron dos técnicas de entrenamiento: en la primera se usaron los dígitos aislados con solo dos repeticiones; en la segunda se combinaron dígitos embebidos en las cadenas con dígitos aislados, obviamente para ambas codificaciones el comportamiento del segundo entrenamiento fue mejor.

En la tabla 6.6 se muestran los resultados usando el primer entrenamiento y los parámetros reducidos de la KLT. Si bien la precisión no es muy alta, debe considerarse que la base de entrenamiento es muy pequeña y los dígitos aislados no es la mejor opción como entrenamiento.

Para los experimentos con la KLT, se aplicaron segmentaciones acústicas basadas en el método MLR, donde hubo que realizar modificaciones al algoritmo para ajustar los cambios espectrales con valores distintos a los obtenidos para palabras aisladas.

longitud de la cadena en dígitos	Número de repeticiones	errores 0	1	2	3	4	5	6	7
2	156	137	16	3	0	0	0	0	0
3	148	110	24	10	4	0	0	0	0
4	153	111	30	7	4	1	0	0	0
5	160	105	38	11	5	1	0	0	0
7	153	84	45	14	7	2	1	0	0

Tabla 6 8. Matriz de confusión para palabras conectadas en inglés de la base de TI, método DTW de un solo paso, con parámetros KLT, entrenamiento embebido y con palabras aisladas.

En la tabla 6.7 se muestran los resultados con el mismo entrenamiento y los coeficientes LPC, y la precisión es muy semejante al caso anterior. En ambas tablas puede observarse que el número de repeticiones utilizado en cada longitud de cadena varía, ya que de 160 repeticiones disponibles algunas de éstas estaban dañadas .

longitud de la cadena en dígitos	Número de repeticiones	errores en % 0	1	2	3	4	5	6	7
2	156	87.8	10.3	1.9	0	0	0	0	0
3	148	74.3	16.2	6.8	2.7	0	0	0	0
4	153	72.5	19 6	4.6	2.6	0.7	0	0	0
5	160	65.6	23 8	6.9	3.1	0.6	0	0	0
7	153	54.9	29.4	9.1	4 6	1 3	0.7	0	0

Tabla 6 9. Matriz de confusión en % para palabras conectadas en inglés de la base de TI, método DTW de un solo paso, con parámetros KLT, entrenamiento embebido y con palabras aisladas

En la tabla 6.8 se muestran los resultados usando un entrenamiento combinado entre palabras embebidas y palabras aisladas, en primer término se duplicó el número de repeticiones del entrenamiento, que aunado al uso de palabras embebidas fue crucial para mejorar el rendimiento. En esta tabla se usaron los parámetros reducidos de la KLT.

Los resultados mostrados en la tabla 6.7 se muestran ahora en porcentajes en la tabla 6.9.

Para los coeficientes LPC se obtuvieron los resultados mostrados en la tabla 6.10 con el entrenamiento embebido en combinación con palabras aisladas. De esta tabla puede observarse que los resultados son muy similares a los obtenidos por la KLT. Sin embargo, se observa una menor precisión a la obtenida por los autores presentados en el subtema anterior. Se atribuye esta baja en el rendimiento a que la base de entrenamiento utilizada en este trabajo es mucho menor a la utilizada por los autores mencionados.

longitud de la cadena en dígitos	Número de repeticiones	errores 0	1	2	3	4	5	6	7
2	156	135	20	1	0	0	0	0	0
3	148	118	25	5	0	0	0	0	0
4	153	114	28	8	3	0	0	0	0
5	160	110	35	11	4	0	0	0	0
7	153	72	52	18	8	3	0	0	0

Tabla 6.10. Matriz de confusión para palabras conectadas en inglés de la base de TI, método DTW de un solo paso, con coeficientes LPC, entrenamiento embebido y con palabras aisladas.

Nuevamente se presentan los resultados de la tabla 6.10 en porcentajes para una mayor claridad en la tabla 6.11.

longitud de la cadena en dígitos	Número de repeticiones	errores en % 0	1	2	3	4	5	6	7
2	156	86.5	12.8	0.7	0	0	0	0	0
3	148	79.7	16.9	3.4	0	0	0	0	0
4	153	74.5	18.3	5.2	2.0	0	0	0	0
5	160	68.7	21.9	6.9	2.5	0	0	0	0
7	153	47.0	34.0	11.8	5.2	2.0	0	0	0

Tabla 6.11. Matriz de confusión en % para palabras conectadas en inglés de la base de TI, método DTW de un solo paso, con coeficientes LPC, entrenamiento embebido y con palabras aisladas.

CONCLUSIONES

La presentación de la transformada de Karhunen-Loeve como codificador de voz ha generado entre los expertos indiferencia y escepticismo, lo cual es hasta cierto punto explicable, ya que la KLT es bien conocida por ser una transformada ortogonal óptima pero no rápida, que inclusive se usa como punto de referencia para obtener transformadas semióptimas pero más rápidas.

En estas consideraciones, se ha pasado por alto que para la codificación o reconocimiento de voz es necesario realizar una postcodificación a partir de la correlación en tiempo corto para obtener los parámetros por trama, como los multicitados LPC o cepstral. El tiempo de procesamiento para obtener estos últimos en una subpalabra es mayor al tiempo para obtener los parámetros reducidos KLT para reconocimiento, que son los primeros vectores y valores propios. En consecuencia, que la KLT no sea un algoritmo rápido no tiene efectos significativos para los sistemas de reconocimiento propuestos.

Resulta irónico que si bien los parámetros KLT determinan que no sea ésta una transformada rápida, son precisamente estos los que proporcionan la capacidad para ubicar subpalabras en clases separables, lo que no ocurre con los parámetros de transformadas rápidas conocidas.

Aunado a lo anterior, la utilización de unidades de procesamiento por subpalabra y no por trama determina que la clasificación de palabras sea de 1.3 a 1.5 veces más rápida que si se usan coeficientes LPC o cepstral, para los diferentes sistemas de reconocimiento de palabras aisladas, y con una precisión semejante. Estos resultados constituyen la aportación toral del presente trabajo.

Si bien las precisiones obtenidas para palabras aisladas sobrepasan el 99%, se puede aún elevar su precisión aumentando el número de subpalabras, pero entonces requeriría mayor tiempo de procesamiento.

Otro de los aspectos planteados en el trabajo y que resultan innovadores en el área, es la utilización de segmentos acústicos. En gran parte de los sistemas actuales de reconocimiento, se usan unidades fonéticas como segmentos. El objetivo de utilizar subpalabras acústicas fue

el de obtener segmentos relevantes analítica y lingüísticamente, sin las complicaciones que se tienen para segmentar automáticamente unidades lingüísticas.

De lo descrito anteriormente, se desprende la ventaja de utilizar parámetros mayores a los de una trama, y donde la KLT tiene un papel relevante, y que coincide con la actual dirección en las investigaciones sobre el reconocimiento de palabras continuas.

Otra de las aportaciones que maneja el sistema de codificación utilizado es el de decimar las magnitudes de la transformada corta de Fourier en bandas críticas. De hecho antes de codificar por la KLT, los vectores de bandas críticas proporcionan una decimación perceptual de la trama semejante a los coeficientes LPC o cepstral, del doble de dimensión pero mucho más rápidas de obtener. En virtud de que la rapidez de procesamiento es crucial en aplicaciones comerciales, resulta también en una codificación muy útil.

Si bien en los sistemas DTW y VQ utilizados en palabras aisladas se obtienen precisiones muy altas, esto no es así para los sistema HMM. Como se ha descrito, los sistemas HMM no modelan con precisión estados tan amplios como las subpalabras. De hecho, en la mayoría de los sistemas utilizados se asocia todo un modelo HMM por fonema, haciendo el procesamiento muy intensivo.

En palabras conectadas, los resultados son buenos considerando la base tan pequeña de entrenamiento en los casos presentados. La extensión del DTW para palabras aisladas es sencilla y natural haciendo uso del método de un solo paso, y confirma la pertinencia del uso de la KLT.

Los algoritmos fueron finalmente transportados en una computadora personal de mediana capacidad, y en esta plataforma fueron hechas las mediciones de velocidad. En lo futuro es contar con paquetes de comandos aplicados a sistemas operativos y comerciales.

De los experimentos con la base de datos en español, la alta precisión obtenida confirma las condiciones propias de nuestro idioma que facilitan el reconocimiento comparado con otros idiomas. Los resultados aumentan el interés por realizarlos para distintas condiciones como uso de fonemas y para palabras continuas, así como el desarrollo de sistemas voz-texto-voz.

APÉNDICE. ARTÍCULOS PUBLICADOS

Resultados parciales de la tesis han sido enviados a diferentes congresos. A continuación se mencionan algunos y en que memorias han sido publicados. Cabe mencionar que lo publicado no agota la totalidad de los resultados obtenidos, otros resultados y en especial de HMM y de palabras conectados están en preparación para su publicación. Destaca el artículo 11 que fue nominado para mejor artículo dentro del congreso internacional *10th Intenational Conference on Signal Processing Applications & Technology*, donde se presentaron más de 400 artículos.

1. A. Herrera, V.R. Algazi, K. Brown y D. Irvine. "Subword Segmentation Alternatives for isolated and Connected Words Recognition". *Proceedings de VIII European Signal Processing Conference*, vol. I, pp. 107-110, 1994.

2. A. Herrera, V.R. Algazi, y D. Irvine. "An Acoustic Approach for Isolated Word Recognition". *Proceedings de The International Conference on Signal Processing Applications and Technology*, vol. II, pp. 1677-1681, 1994.

3. A. Herrera, V.R. Algazi, y A. Mondragón. "An Isolated Speech Recognition Approach Using the KLT". *Proceedings de 1995 IEEE Workshop on Automatic Speech Recognition*, pp. 191-192, 1995.

4. A. Herrera, A. Ramos y K. Yamasaki. "Speech Detection in High Noise Conditions". *Proceedings of the 7th Intenational Conference on Signal Processing Applications & Technology*, vol. I, pp. 1774-1778, 1996.

5. A.F. Mondragón y A. Herrera. "Speech Recognition Techniques using Acoustic Segmentation". *Proceedings of the IASTED International Conference on Signal and Image Processing*, pp. 26-29, 1996.

6. Abel Herrera, "Reconocimiento automático de voz utilizando la transformada de Karhunen-Loève". *Memorias del Simposio La Investigación en la Facultad de Ingeniería 1996*, Facultad de Ingeniería, UNAM, 1997, resumen

7. Rafael Sánchez y Abel Herrera. "Reconocimiento adaptable de palabras utilizando redes neuronales". *Memorias del Simposio La Investigación en la Facultad de Ingeniería 1997*, Facultad de Ingeniería, UNAM, 1997, resumen .

8. A. Herrera, M. Martinez, V.R. Algazi y O. Sánchez, "An Acoustic Isolated Speech Recognition, Approach Using the KLT and VQ". *The Proceedings of the 8th International Conference on Signal Processing Applications & Technology*, vol. II, pp. 1739-1742, 1997.

9. Alexandre Bouchet y Abel Herrera. "Reconocimiento Automático de Palabras Aisladas y Conectadas Usando DTW". *Memorias del Simposio La Investigación en la Facultad de Ingeniería 1998*, Facultad de Ingeniería, UNAM, (próximas a aparecer).

10. A. Herrera, R. Ibarra y V.R. Algazi. "An Isolated Speech Recognition Method Using KLT". *Proceedings of the IASTED International Conference on Signal and Image Processing*, pp. 26-29, 1999.

11. A. Herrera, A. Gardida y V.R. Algazi. "A VQ Isolated Speech Recognition Acoustic Method Using the KLT". *Proceedings of the 10th International Conference on Signal Processing Applications & Technology*, diskette 1, 1999.

12. A. Herrera, y A. Gardida. "A Multisection VQ Isolated Speech Recognition Method Using the KLT". *Proceedings of the International Conference on Communications 2000*, pp. 1133-1136, 2000.

13. A. Herrera, R. Sánchez y B. Psenicka. "Adaptive Isolated Word Recognition Using a New NN, and KLT". *Proceedings of the IASTED International Conference on Signal and Image Processing*, pp. 320-323, 2000.

REFERENCIAS

[1] R.K. Potter, G.A. Kopp y H. Green. *Visible Speech*. 1ª reedición, Dover, New York, 1966.

[2] K.H. Davis, R. Biddulph, y S. Balashek. "Automatic Recognition of Spoken Digits". *Journal Acoustic Society of America*, vol. 24(6), 1952, pp. 637-642.

[3] H. Dudley y S. Balashek. "Automatic Recognition of Phonetic Patterns in Speech". *Journal Acoustic Society of America*, vol. 30, 1958, pp. 721-739.

[4] P.B. Denes. "Automatic Speech Recognition: Old and New Ideas". *Proceedings of the IEEE Symposium on Speech Recognition*, Academic Press, 1975, pp. 73-82.

[5] E.P. Neuburg. "Philosophies of Speech Recognition". *Proceedings of the IEEE Symposium on Speech Recognition*, Academic Press, 1975, pp. 83-95.

[6] T.B. Martin. "Applications of Limited Vocabulary Recognition Systems". *Proceedings of the IEEE Symposium on Speech Recognition*, Academic Press, 1975, pp. 55-73.

[7] F. Itakura. "Minimum Prediction Residual Applied to Speech Recognition ". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23(1), 1975, pp. 67-72.

[8] D. Ray. Reddy. "Speech Recognition by Machine: A Review". *Proceedings of the IEEE*, vol. 64(4), 1976, pp. 502-531.

[9] J. Mariani. "Recent Advances in Speech Processing". *Proceedings of ICASSP'89*, vol. 1, 1989, pp.429-440.

[10] L.R. Rabiner, S.E. Levinson, A.E. Rosenberg y J.G. Wilpon. "Speaker-Independent Recognition of Isolated Words Using Clustering Techniques". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27(4), 1979, pp.336-349.

[11] A. Buzo, H. Martínez, y C Rivera. "Discrete Utterance Recognition Based Upon Source Coding Techniques". *Proceedings of ICASSP'82*, vol. 1, París, 1982, pp.539-542.

[12] F. Jelinek. "The Development of an Experimental Discrete Dictation Recognizer". *Proceedings of the IEEE*, vol. 73(11), 1985, pp. 1616-1624

[13] L.R. Rabiner. "A Tutorial of Hidden Markov Models and Selected Applications in Speech Recognition". *Proceedings of the IEEE*, vol.77(2), 1989, pp. 257-286.

[14] J. Savage y A Herrera. "Isolated-Word Speech Recognition Using Hidden Markov Models and Multisection Vector Quantization Code Books". *Proceedings of the International Symposium on Communications Theory & Applications*, Moscú, 1991, sin páginas.

[15] R.W. Christiansen y C.K. Rushforth. "Detecting and Locating Key Words in Continuous Speech Using Linear Predictive Coding". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25(2), 1981, pp.84-297.

[16] S. Boll, J. Porter y L. Bahler. "Robust Syntax Free Speech Recognition". *Proceedings of ICASSP'88*, vol 1, 1988, pp.170-182.

[17] J.G. Wilpon, C.H Lee y L.R. Rabiner "Applications of Hidden Markov Models for Recognition of a Limited Set of Words in Unconstrained Speech". *Proceedings of ICASSP'89*, vol. 1, 1989, pp.254-257.

[18] H. Sakoe y S. Chiba. "A Dynamic Programming Approach to Continuous Speech Recognition". *Proceedings of the International Congress on Acoustics 1971*, Budapest, Hungria, Rep. 20-C-13, 1971.

[19] H. Sakoe. "Two Level DP-Matching - A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition," en *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-27, No.6, Diciembre 1979, pp 588-595.

[20] C.S. Myers y L.R. Rabiner. "A Level Building Time Warping Algorithm for Connected Word Recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29(2), 1981, pp. 284-297.

[21] C.H. Lee y L.R. Rabiner. "A Frame Synchronous Network Search Algorithm for Connected Word Recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37(11), 1989, pp. 1649-1658.

[22] H. Ney. "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, No.2, abril 1984, pp. 263-271.

[23] T.K. Vintsyuk. "Element-Wise Recognition of Continuous Speech Composed of Words from a Specified Dictionary". *Kibernetica*, vol. 7, Marzo-Abril, 1971, pp. 133-143.

[24] J.S. Bridle y M.D. Brown. "Connected Word Recognition Using Whole Word Templates". *Proceedings of the International Acoustic Autumn Conference*, noviembre 1979, pp.25-28.

[25] L.R. Rabiner y J.G. Wilpon, "A simpliifed, Robust Training Procedure for Connected Digit Recognition". *Journal of the Acoustic Society of America*, vol. 68, No.5, noviembre 1980, pp. 1271-1276.

[26] C.S. Myers y L.R. Rabiner. "Connected Digit Recognition Using A Level Building DTW Algorithm". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29(3), 1981, pp. 351-363.

[27] L.R. Rabiner, A. Bergh and J.G. Wilpon. "An Improved Training Procedure for Connected Digit Recognition". *The Bell System Technical Journal*, vol. 61, No. 6, Julio-Agosto 1982, pp. 981-1001.

[28] L.R. Rabiner, J.G. Wilpon, A.M. Quinn y S.G. Terrace. "On the Application of Embedded Digit Training to Speaker Independent, Connected Digit Recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32(2), 1984, pp. 272-280.

[29] L.R. Rabiner, C.H. Lee. J. G. Wilpon y B.H. Juang. "A Segmental K-means Training Procedure for Connected Word Recognition". *AT&T Bell Technical Journal*, vol. 61, No. 3, Mayo-Junio 1986, pp. 21-40.

[30] L.R. Rabiner, J. G. Wilpon y B.H. Juang. "Performance Evaluation of a Connected Digit Recognizer". *Proceedings of ICASSP'87*, vol. 1, Dallas, 1986, pp. 101-104.

[31] L.R. Rabiner, C.H. Lee. J. G. Wilpon y B.H. Juang. "HMM Clustering for Connected Word Recognition". *Proceedings of ICASSP'89*, vol. 1, Glaswog, 1989, pp. 405-408.

[32] L.R. Rabiner y B.H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, 1993.

[33] K.F. Lee, H.W. Hon y D.R Reddy. "An Overview of the SPHINX Speech Recognition System". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, 1990, pp. 600-610.

[34] Y.L. Chow, M.O. Dunham, O.A. Kimball, M.A. Krasner, y G.F. Kubala. "BBYLOS: The Continuous Speech Recognition System". *Proceedings of ICASSP'87*, vol. 1, 1987, pp. 89-92.

[35] D.B. Paul. "The Lincoln Robust Continuous Speech Recognizers". *Proceedings of ICASSP'89*, vol. 1, Glaswog, 1989, pp. 449-452.

[36] M. Weintraub. "Linguistic Constraints in Hidden Markov Models Based Speech Recognition". *Proceedings of ICASSP'89*, vol. 1, Glaswog, 1989, pp. 699-702.

[37] The Speech Technical Committee, editor B.H. Juang. "The Past, Present and Future of Speech Processing". *The IEEE Signal Processing Magazine,* vol. 15(3), 1998, pp. 24-48.

[38] C.H. Lee, L.R. Rabiner, R. Pieraccini y J.G. Wilpon."Acoustic Modeling for Large Vocabulary Speech Recognition". *Computer Speech and Language,* vol. 37(11), 1989, pp. 1649-1658.

[39] J.A. Markowitz. *Using Speech Recognition.* Prentice Hall, Upper Saddle River, 1995.

[40] A. Lavie, A. Waibel, L.Levin, M. Finke, D. Gates, M. Gravalda, T. Zeppenfield y P. Zhan. "Janus III: Speech-to-Speech Translation in Multiple Languages". *Proceedings of ICASSP'97*, vol. 1, Munich, 1997, pp. 99-102.

[41] E. Vidal. "Finite State Speech-to-Speech Translation". *Proceedings of ICASSP'97*, vol. 1, Munich, 1997, pp. 111-114.

[42] L.R. Rabiner y R.W. Schafer. *Digital Processing of Speech Signals.* Prentice Hall, Englewood Cliffs, 1978.

[43] J.D. Markel y A.H. Gray. *Linear Prediction of Speech.* Springer-Verlag, New York, 1976.

[44] V. Algazi, S. Chung, M. Ready, y K. Brown. "Robust LPC Analysis and Synthesis Using the KL Transformation of Acoustic Subwords Spectra". *Proceedings of ICASSP'89*, vol. 1, 1989, pp. 468-471.

[45] V.R. Algazi, D.H. Irvine, K.L. Brown, M.J. Ready, C.L. Caldwell y S. Chung. "Transform Representation of the Spectra of Acoustic Speech Segments with Aplications-II: Speech Analysis, Synthesis, and Coding". *IEEE Transactions on Speech and Audio Processing*, vol. 1(3), 1993, pp. 277-286.

[46] V.R. Algazi, K.L. Brown, M.J. Ready, D.H. Irvine, C.L. Caldwell y S. Chung. "Transform Representation of the Spectra of Acoustic Speech Segments with Aplications-I: General Approach and Application to Speech REcognition". *IEEE Transactions on Speech and Audio Processing,* vol 1(2), 1993, pp. 180-195

[47] Bertram Scharf "Chapter: Critical Bands". *Foundations of Modern Auditory Theory*, editor Jerry V.,Tobias, Academic Press,vol. I, pp. 157-202, New York, 1970.

[48] Lloyd A. Jeffres. "Chapter: Masking". *Foundations of Modern Auditory Theory*, editor Jerry V. Tobias, Academic Press, vol. I, pp. 87-114, New York, 1970.

[49] Harvey Fletcher *Speech and Hearing,* Van Nostrand, New York, 1929.

[50] E. Zwicker and H Fastl. *Psychoacoustics.* Springer-Verlag, Berlin, 1990.

[51] N.R. French and J.C. Steinberg. "Factors Goberning the Intelligibility of Speech Sounds". *J. Acoust. Soc. Amer.*, vol. 19(1), pp 90-119, 1947.

[52] S. Bolaños. *Manual de fonética española para estudiantes extranjeros.* UNAM, México, 1975.

[53] S.A. Williams, *An Acoustic Analysis of the Spanish Sound System.* Tesis doctoral, Georgetown University, 1982.

[54] A. Quilis *Fonética acústica de la lengua española.* Gredos, Madrid, 1981.

[55] A Bolaño e Isla Breve *manual de fonética elemental,* 2ª edición, Porrua, México, 1968.

[56] Real Academia Española. *Ortografía de la lengua española.* Espasa, Madrid, 1999.

[57] Real Academia Española *Diccionario de la Lengua Española.* 19ª edición, Espasa-Calpe, , Madrid, 1970.

[58] J.G. Moreno de Alba Estructura de la Lengua Española. Trillas, 2ª edición, México, 1992.

[59] H.L. Van Trees. *Detection, Estimation, and Modulation Theory.* Wiley, New York, 1968.

[60] J.L. Melsa y D L Cohn. *Decision and Estimation Theory.* Parte I, Mc.Graw-Hill, Tokyo, 1978.

[61] K. Fukunaga. *Statistical Pattern Recognition.* 2ª. edición, Academic Press, San Diego, 1990.

[62] L.R. Rabiner, y M.R. Sambur. "An Algorithm for Determining the Endpoints of Isolated Utterances" *The Bell System Technichal Journal*, vol. 54, No. 2, pp.297-315, febrero 1975.

[63] L.F. Lamel, L.R. Rabiner, L.R. Rosenber, y J.G. Wilpon. "An Improved Endpoint Detector for Isolated Word Recognition". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, No. 4, agosto 1981.

[64] A. Ramos y K. Yamasaki. Detección de principio y fin de señales de voz. Tesis de licenciatura en Ingeniería, UNAM, 1996.

[65] N. Ahmed y K.R.Rao *Orthogonal Transforms for Digital Signal Processing.* Springer-Verlag, Berlin-Heidelberg, 1975.

[66] E. Oja. *Subspace Methods for Pattern Recognition.* John Wiley & Sons Inc., New York, 1983.

[67] A. K. Jain *Fundamentals of Digital Image Processing* Edit. Prentice Hall, Englewood Cliffs, 1989.

[68] H. Stark y J.W. Woods. *Probability, Random Processes, and Estimation Theory for Engineers.*2ª. edición, Prentice Hall, Englewood Cliffs, 1986.

[69] M. Fiedler *Special Matrices and their Applications in Numerical Mathematics.* Martinus Nijhoff Publishers, Praga, 1986.

[70] C.S. Chen y K.S. Huo. "Karhunen-Loeve Method for Data Compression and Speech Synthesis". *IEEE Proceedings*, vol 138(5), 1991, pp. 377-380.

[71] T.W. Parsons.*Voice and Speech Processing* . McGraw-Hill, New York, 1987.

[72] A. Herrera, V.R. Algazi, K. Brown and D. Irvine. "Subword Segmentation Alternatives for Isolated and Connected Words Recognition". *Proceedings of The VIII European Signal Processing Conference*, vol. I, 1994, pp. 107-110.

[73] R. M. Gray y A. Gersho. *Vector Quantization and Signal Compressing.* Kluwer Academic, 1992.

[74] G. Ford. *Notas del curso Pattern Recognition.* University of California, Davis. 1993.

[75] A. Buzo, A.H. Gray, R.M. Gray y J.D. Markel. "Speech Coding Based upon Vector Quantization". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28(5), 1980, pp. 562-574.

[76] J.T. Tou y R.C. González. *Pattern Recognition Principles.* 4ª impresión, Addison-Wesley Publishing Company, Reading, 1982.

[77] B.H. Juang y L.R. Rabiner."The Segmental K-means Algorithm for Estimating Parameters of Hidden Markov Models". *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 38(9), 1990, pp. 1639-1641.

[78] J. Picone. "Continuous Speech Recognition Using Hidden Markov Models". *IEEE Acoustics, Speech and Signal Processing Magazine,* julio 1990, pp. 26-41.

[79] Y.Q. Gao, Y.B. Chen y T.Y. Huang. "A New Method for Estimation of Hidden Markov Models Parameters". *Proceedings of International Conference on Pattern Recognition*, vol. 2, junio 1990, pp. 27-30.

[80] L.R. Bahl, P.F. Brown, P.V. De Souza y R.L.Mercer. "A New Algorithm for the Estimation of Hidden Markov Models". *Proceedings of ICASSP'88*, vol. 1, New York, 1988, pp. 493-496.

[81] R. Bahl, P.F. Brown, P.V. De Souza y R.L.Mercer. "Speech Recognition with Continuous-Parameter Hidden Markov Models". *Computer, Speech and Language*, vol. 2(3-4), 1987, pp. 219-234.

[82] Y. Ephraim y L.R.Rabiner. "On the Relations between Modeling Approaches for Information Sources". *Proceedings of ICASSP'88*, vol. 1, New York, 1988, pp. 24-27.

[83] A. Kriouile, J.F. Mari y J.P.Haton. "Some Improvements in Speech Recognition Algorithms Based on HMM". *Proceedings of ICASSP'90*, vol. 1, Albuquerque, 1990, pp. 545-548.

[84] L.R. Rabiner y S.E. Levinson. "Isolated and Connected Word Recognition - Theory and Selected Applications". *IEEE Transactions on Communications*, vol. Com-29, May 1982,pp 621-659.

[85] L.R. Rabiner, C.H. Lee. J. G. Wilpon y B.H. Juang. "A Model-Based Connected-Digit Recognition System Using either Hidden Markov Models or Templates". *Computer, Speech and Language*, n_o 1, Diciembre 1986, pp.167-197.

[86] L.R. Rabiner y M.R. Sambur. "Experiments in the Recognition of Connected Digits". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-24, No. 2, Abril 1976 pp. 170-182.

[87] H. Iwamida, S. Katagari, E. McDermott e Y. Tohkura, "A Hybrid Speech Recognition System Using HMMs with an LVQ-Trained Algorithm". *Proceedings of ICASSP'90*, vol. 1, Albuquerque, Abril 1990, pp. 489-492.

[88] M. Falkhausen, S.A. Euler y D. Wolf, "Improved Training and Recognition Algorithms with VQ_Based Hidden Markov Models". *Proceedings of ICASSP'90*, vol.1, Albuquerque, Abril 1990, pp. 549-552

[89] L. Niles y H. Silverman. "Neural Networks, Maximum Mutual Information Training and Maximum Likelihood Training". *Proccedings of ICASSP'90*, vol.1, Albuquerque, Abril 1990, pp. 493-496.

[90] Y. Normandin y S. Morgera. "An Improved MME Training Algorithm for Speaker-Independent, Small Vocabulary, Continuous Speech Recognition". *Proceedings of ICASSP'91*, vol.1. Toronto, Mayo 1991, pp. 537-540.

[91] Y. Gao y J. Xie. "A Model Block-Training Method for HMM-Based Speech Recognition Systems". *Proceedings of ICASSP'90*, vol.1, Albuquerque, Abril 1990, pp. 541-544.