UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

EVALUACIÓN DE PRONUNCIACIÓN POR TONO PARA ENSEÑANZA DE SEGUNDO IDIOMA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELECTRICISTA

JUAN PABLO ARIAS APARICIO

Profesor Guía:
Néstor Becerra Yoma

Miembros de la Comisión:
Nelson Baloian Tataryan
Jorge Wuth Sepúlveda

Santiago de Chile
Noviembre de 2008

Resumen de la Memoria
para optar al título de
Ingeniero Civil Electricista
Por: Juan Pablo Arias Aparicio
Fecha: 30/10/2008
Prof. Guía: Sr. Néstor Becerra Yoma

"Evaluación de Pronunciación por Tono para Enseñanza de Segundo Idioma"

La prosodia es un elemento fundamental en el proceso de enseñanza de una lengua extranjera, ya que provee al hablante de características esenciales en la comunicación como naturalidad y fluidez. En virtud de lo anterior, un sistema de enseñanza de segundo idioma asistido por computador debe poseer un módulo mediante el cual los estudiantes puedan entrenar su percepción y producción de prosodia.

El sistema de evaluación de pronunciación por tono que se describe y desarrolla a lo largo de este trabajo pretende evaluar la entonación de un estudiante mediante una comparación entre su propia voz y una señal pregrabada de referencia. El usuario escucha una cierta palabra u oración con una determinada melodía, y luego intenta imitarla. Finalmente, el sistema entrega un puntaje o nota conforme a la similitud alcanzada. La implementación de la herramienta involucra el uso del algoritmo de alineamiento temporal dinámico (DTW, *Dynamic Time Warping*) y la estimación de la frecuencia fundamental f_0.

Se realizaron diversos experimentos utilizando una base de datos compuesta por palabras y oraciones en inglés, grabada por locutores nativos del español. Para la mejor configuración, el coeficiente de correlación entre los resultados entregados por el sistema y la evaluación esperada es igual a 0,87. Por otra parte, para evaluación de acentuación se obtiene una tasa de error igual a 20,97 % en el mejor de los casos.

Agradecimientos

Quisiera expresar mi profundo agradecimiento a mis padres Isabel y Raúl por el apoyo y cariño incondicional que me han entregado durante toda mi vida. Gracias por la infinita confianza; por inspirar en mí el esfuerzo y la perseverancia; y por estar siempre tan cerca. A mi hermano Miguel por su inagotable alegría; por quererme tanto y enseñarme a ver la vida desde otra perspectiva. A mi polola Natalia por estar junto a mi durante esta útlima etapa, por su dedicación, por ayudarme a superar las dificultades y por ser un motivo para enfrentar cada día con entusiasmo.

A mis queridos amigos de SLB, compañeros institutanos que han sido parte de momentos inolvidables de mi vida, les quiero agradecer por compartir conmigo la pasión por el conocimiento y la música. A mis compañeros de universidad Javier y Leo por su cercanía durante todos estos años de estudio y por su sincera amistad. Le quiero dar gracias también a mi primo Marcelo, por estar conmigo prácticamente toda mi vida y por entregarme su valiosa ayuda que muchas veces no pude retribuir.

Agradezo a mi profesor Néstor Becerra por el conocimiento que me ha transmitido y por haberme dado la oportunidad de desarrollar este interesante tema de investigación, y a mis compañeros del LPTV por compartir conmigo su experiencia para realizar este trabajo.

... Dedicado a mis padres

Isabel Aparicio y Raúl Arias

Índice general

Índice de figuras

Índice de Tablas

Glosario

ASR: *Automatic Speech Recognition*, reconocimiento automático de voz.

Acento: Énfasis que se imprime a una sílaba distinguiéndola del resto de la palabra.

Alineamiento: Proceso que consiste en asociar un vector de parámetros acústicos de una señal de voz con otra.

CALL: *Computer-Aided Language Learning*, enseñanza de segundo idioma asistida por computador.

Características suprasegmentales: Características de la voz en un nivel superior a los segmentos fonéticos, como la entonación, la acentuación, la duración y el ritmo.

Coeficientes cepstrales: Parámetros acústicos que caracterizan la información espectral de un segmento de voz.

DCT: *Discrete Cosine Transform*, transformada discreta coseno.

DTW: *Dynamic Time Warping*, alineamiento temporal dinámico.

EER: *Equal Error Rate*.

Entonación: Modulación de la voz en la secuencia de sonidos del habla que puede reflejar diferencias de sentido, de intención, de emoción y de origen del hablante.

FA: Falsa aceptación.

FR: Falso rechazo.

Fonología: Rama de la lingüística que estudia los elementos fónicos, atendiendo a su valor distintivo y funcional.

Fonética: Estudio acerca de los sonidos de uno o varios idiomas, sea en su fisiología y acústica, sea en su evolución histórica.

Frame: Segmento de voz de una determinada duración, resultado del proceso de enventanado. Unidad mínima de análisis.

Frecuencia fundamental f_0: Frecuencia más baja de la descomposición armónica de una señal.

Lenguaje Natural: Situación que se presenta en un diálogo conversacional cuando el usuario de un sistema expresa una solicitud utilizando más palabras de las requeridas.

MFCC: *Mel-frequency Cepstral Coefficients*, coeficientes cepstrales en la escala de Mel.

Modulación: Variar con fines armónicos las cualidades del sonido en el habla o en el canto.

Pitch: Percepción de la frecuencia fundamental f_0 por el ser humano.

Pragmática: Disciplina que estudia el lenguaje en su relación con los usuarios y las circunstancias de la comunicación.

Prosodia: Disciplina que estudia formalmente los elementos de la expresión oral como acentos, entonación y duración.

ROC: Reciever Operating Charasteristic.

Sonido sonoro: Sonido que involucra vibración de las cuerdas vocales. También se denominan *tonales*.

Sonido áfono: Aquel sonido que no se genera a partir de la exitación de las cuerdas vocales. También se denominan *sordos*.

Tono: Grado de elevación del sonido o de la voz.

Capítulo 1

Introducción

1.1. Motivación

Ciertamente, la enseñanza de segundo idioma asistida por computador (CALL, *Computer-aided language learning*) ofrece enormes ventajas a los estudiantes como una herramienta complementaria a la instrucción presencial. La interactividad y el entretenimiento implícito en el *software* educativo hacen más efectivo el proceso de aprendizaje, ya que incrementan la motivación y permiten desarrollar actividades y ejercicios sin necesidad de la constante supervisión de un profesor. Muchas veces, los estudiantes que recién comienzan el proceso de adquisición de un idioma extranjero sienten timidez e incomodidad al pronunciar palabras en otra lengua, por lo que interactuar con un computador puede ayudar significativamente a los alumnos retraídos.

En los últimos años, CALL ha experimentado grandes cambios gracias al desarrollo de la ciencia y la tecnología. Los sistemas de instrucción tradicionales basados en texto e imágenes estáticas han sido reemplazados por interacciones cada vez más parecidas a la vida real, que involucran diálogos entre el estudiante y el computador. En esta evolución de CALL, tecnologías como el reconocimiento automático de voz (ASR, *au-*

tomatic speech recognition); la síntesis de voz (TTS, *text to speech*); y el procesamiento de señales han jugado un rol fundamental.

La televisión y la Internet diariamente exponen a las personas a textos y diálogos en otros idiomas. Por esta razón, la mayoría de los estudiantes poseen grandes habilidades receptivas, ya que comprenden sin dificultades el inglés hablado y escrito. Sin embargo, sus habilidades productivas son de muy baja calidad. Esto se debe a que los estudiantes en su vida cotidiana no poseen actividades conversacionales, lo que se traduce en problemas de pronunciación y baja capacidad para expresar ideas con claridad.

Estudiar una lengua extranjera involucra la adquisición de diferentes aspectos del lenguaje tanto escrito como hablado, siendo la prosodia uno de los más importantes. La prosodia se refiere a la modulación de las características acústicas del habla en un nivel superior a los segmentos fonéticos, la cual contiene información lingüística (relativa al lenguaje y la comunicación) y paralingüística (emociones, actitudes e intenciones).

1.2. Objetivos

En este trabajo, se propone un método de evaluación de entonación para la enseñanza de segundo idioma, el cual extrae y compara las características prosódicas de una señal de referencia y una de prueba. Para ello, el estudiante escucha la señal pregrabada y luego graba su propia voz tratando de imitar la melodía. Una vez que el sistema procesa la señal, entrega al alumno un puntaje o nota conforme al nivel de similitud alcanzado. Adicionalmente, se extiende la funcionalidad del método para resolver el problema de evaluación de acentuación o *stress*, para lo cual se agrega la energía de la señal como parámetro a la comparación. En este caso, el sistema indica

al usuario si su acento es igual al de referencia o no.

Es muy difícil establecer modelos de entonación que sean generalizables. Por ejemplo en inglés, se utilizan entonaciones ascendentes para las preguntas cuya respuesta es sí o no, y descendentes para otro tipo de respuestas. Sin embargo, lo anterior puede cambiar radicalmente en función de factores como el contexto, el estado de ánimo y el origen de los hablantes, entre otros. Por esta razón, en este trabajo no se utilizan modelos ni patrones de entonación, sino más bien se entrega al alumno un determinado contexto y una referencia para imitar.

Para implementar y probar el sistema, se deben concretar los siguientes objetivos específicos:

- Analizar el problema de la evaluación de características prosódicas en CALL.
- Proponer un método de evaluación de entonación y acentuación.
- Implementar un método de alineamiento de señales mediante parámetros acústicos, y probar empíricamente un desempeño óptimo.
- Implementar módulos de extracción de características prosódicas.
- Evaluar el desempeño del sistema, tanto para entonación como acentuación.
- Probar el comportamiento del sistema frente a situaciones adversas como pronunciaciones erróneas y micrófonos de baja calidad.

1.3. Estructura de la memoria

El Capítulo 2 tiene como objetivo interiorizar al lector en el problema de evaluación de entonación. Se pretende entregar una base teórica suficiente para comprender los

conceptos aquí abordados, con la finalidad de que este trabajo pueda ser comprensible incluso para personas que no poseen conocimientos avanzados en el área de la biometría y el reconocimiento de patrones. En primer lugar, se contextualiza el problema central de esta memoria dentro de la enseñanza de segundo idioma asistida por computador. Luego se describen las características prosódicas del habla como la entonación y el acento o *stress*. Posteriormente se revisan algunas técnicas de procesamiento de voz como la extracción de la frecuencia fundamental f_0 y finalmente se analiza en detalle el algoritmo de alineamiento temporal dinámico DTW.

El Capítulo 3 muestra el desarrollo e implementación del sistema de evaluación de entonación y acentuación propuesto en este trabajo, describendo detalladamente cada una de las etapas involucradas. Además, se muestra la operación del método en condiciones reales usando distintas configuraciones, para lo cual se efectúan una serie de experimentos. Por último, en el Capítulo 4 se entregan las conclusiones generales de esta memoria.

Capítulo 2

Tecnologías de voz

2.1. Introducción

El presente capítulo tiene por objeto introducir al lector en el problema de la evaluación de entonación, para lo cual se presentan los conceptos y técnicas de procesamiento que son utilizadas en esta memoria. En primer lugar, se presentan definiciones y conceptos asociados a la enseñanza de segundo idioma asistida por computador (CALL). Después, se analizan las características suprasegmentales del habla como la entonación, el acento y la duración. Posteriormente, se analizan los correlatos acústicos de la prosodia, en particular las diversas técnicas existentes de estimación de la frecuencia fundamental f_0. Finalmente se explica en forma detallada el algoritmo de alineamiento temporal dinámico DTW, una herramienta esencial para la técnica de evaluación de entonación propuesta en este trabajo.

2.2. Enseñanza de segundo idioma asistida por computador

La enseñanza de idioma asistida por computador (CALL *Computer-Aided Language Learning*) se define como el aprendizaje y la instrucción de alguna lengua extranjera, donde los computadores y otros recursos computacionales como la Internet son utilizados para presentar los contenidos en forma interactiva. CALL pretende ser una herramienta complementaria a la enseñanza presencial, y de ninguna forma intenta reemplazar la labor de un educador.

2.2.1. Introducción

El propósito de la enseñanza de segundo idioma es mejorar las cuatro habilidades de un hablante: comprensión auditiva; comprensión lectora; capacidad de escribir y capacidad de hablar. Enseñar y potenciar todas estas destrezas son acciones muy difíciles de llevar a cabo en salas de clases con muchos alumnos. Es absurdo pensar por ejemplo que un profesor puede enseñar y evaluar la pronunciación a 40 alumnos que comparten una misma aula en un tiempo razonable. Por esta razón, la mayoría de los estudiantes en edad escolar desarrollan las habilidades receptivas y *no* las productivas. Un sistema de enseñanza asistido por computador es capaz de entrenar a un estudiante incluyendo estas habilidades e incluso otras de más alto nivel, como la fluidez y naturalidad del habla.

Una de las principales ventajas de CALL es que genera aprendices autónomos con libertad de elección. Esto significa que el estudiante puede concentrarse en los contenidos y actividades que desee, eligiendo aquellos que más se acomodan a su experiencia en la lengua que está practicando, pudiendo ejercitar una actividad tantas veces como

sea necesario. Por otra parte, la instrucción asistida por computador incrementa el interés del estudiante, sobre todo cuando la información es personalizada y se adapta al usuario (Traynor P., 2003). Otra ventaja importante es la motivación, ya que el uso de tecnología hace más interesantes los contenidos. La presencia de imágenes, animaciones, sonido y la integración de reconocimiento de voz generan curiosidad y proveen un *contexto*.

A pesar de todos las beneficios que presenta la enseñanza asistida por compuador no ha estado exenta de críticas. Existen autores que afirman que CALL no posee un marco teórico de diseño bien definido (Chapelle C., 1997). Por ejemplo, una distribución inapropiada de los elementos de una interfaz gráfica podría distorsionar la atención y motivación de un estudiante de segunda lengua. Otra desventaja de CALL es que los profesores de idiomas no tienen la convicción de que un sistema computacional tenga la capacidad de emitir un juicio certero sobre la calidad del habla de un estudiante.

Si bien la tecnología de computadores y software ha evolucionado bastante en los últimos años, las exigencias de CALL también han experimentado un fuerte crecimiento, sobre todo en lo que respecta a la **interactividad**: Un menú de selección múltiple, objetos clickeables o secuencias lineales ya no son suficientes. Frente a este requerimiento, el procesamiento de lenguaje natural (NLP, *Natural Language Processing*) y el procesamiento de voz han jugado un rol protagónico. Así, la comunicación entre la máquina y el usuario está provista de interacciones lingüísticas más parecidas a la vida real.

2.2.2. Historia de CALL.

La enseñanza y aprendizaje de lenguas extranjeras asistidos por computador se inicia en los años 60s, como una relación simbiótica entre la tecnología y la pedagogía. Los sistemas estaban implementados en macrocomputadores o *mainframes* de aquella época, y por tanto estaban confinados a las universidades debido a su gran tamaño y alto costo. Como ejemplo, se puede citar a la Universidad de Illinois, con su proyecto PLATO iniciado en 1960.

Algunos investigadores hablan de tres épocas o eras de CALL, las cuales son: **estructural**; **comunicativa**; e **integrativa** (Warschauer M., 1996). La primera de ellas se ubica cronológicamente entre los 60s y fines de los 70s, donde las actividades se basaban en texto. El alumno leía una pregunta, y entregaba en forma escrita su respuesta al computador, el cual evaluaba la exactitud de ésta, y proveía algún tipo de realimentación visual. Con este esquema, la enseñanza se orientaba a la gramática y traducción. CALL para ese entonces resulta ser novedoso y las máquinas aparecen como un instrumento que estimula a los estudiantes a dar una respuesta.

La era comunicativa se inicia a principios de los 80s y finaliza en los 90s. Coincide con la aparición y masificación de los computadores personales (PC) Las aplicaciones de esta era se caracterizan por estar orientadas a la comunicación, donde se enfatiza la interacción y la enseñanza de la gramática queda implícita. Los contenidos son presentados a través de juegos; entrevistas; encuestas; y juegos de rol o *role plays*.

Por último, la era integrativa comienza en los años 90s, y es la que se desarrolla en la actualidad. Coincide con el desarrollo de las tecnologías multimedia (sonido, animaciones, imágenes y texto) y las redes de computadores (fundamentalmente Internet), así como también el avance de disciplinas como el procesamiento de señales, el

reconocimiento automático y síntesis de voz. Se centra en la interacción social, y los computadores son utilizados para generar un diálogo auténtico, lo más cercano posible a un escenario de interacción social verdadero, para lo cual evidentemente son necesarias interfaces hombre-máquina inteligentes. En la Tabla 2.1 se muestran las características que determinan a cada una de las tres eras de CALL.

Etapa	1970-1980: Estructural	1970-1980: Comunicativa	1990-a la fecha: Integrativa
Tecnología	Macrocomputadores	PC	Multimedia e Internet
Paradigma de enseñanza	Gramática y Traducción	Enseñanza Comunicativa	Basado en Contenido
Visión del Idioma	Estructural	Cognitivo	Socio-cognitivo
Uso principal de Computadores	Ejercicios y práctica	Ejercicios comunicativos	Diálogo auténtico

Tabla 2.1: *Las fases de CALL según Warschauer.*

2.2.3. Tendencias actuales de CALL

Como se ha señalado en la subsección anterior, las tecnologías de voz y el uso de los computadores personales han dado origen a un nuevo paradigma en la enseñanza de segundo idioma. En particular, el reconocimiento automático de voz ha sido exitosamente aplicado a CALL, y mediante su uso se han desarrollado diversos prototipos que utilizan esta tecnología para incentivar a los usuarios a utilizar el lenguaje hablado incluyendo respuestas por voz e incluso conversaciones entre el alumno y el computador.

Con el fin de ayudar a los estudiantes a establecer una asociación entre los sonidos del habla y su escritura, algunos programas educativos han implementado los denominados ejercicios *reading aloud*. La idea es entregar al usuario un texto para que lo

lea en voz alta, mientras un motor de reconocimiento de voz entiende cada una de las palabras que han sido pronunciadas. La mayoría de estos sitemas utilizan un ASR dependiente del locutor con un vocabulario reducido. Este tipo de ejercicios han sido aplicados tanto a enseñanza de segundo idioma como a la alfabetización (Eshani F.& Knodt E., 1998).

Otro tipo de sistemas más avanzados implementan verdaderos diálogos entre el usuario y el computador, los cuales consisten en interacciones lingüísticas que simulan una situación real. Para lograr esto, se hace hablar a los usuarios mediante estímulos gráficos o simplemente a través de una pregunta directa. Existen básicamente dos enfoques de diseño: respuesta cerrada y respuesta abierta. El primero hace referencia a sistemas en los cuales el universo de respuestas posibles es acotado, para lo cual se presentan en pantalla múltiples alternativas de las cuales se debe escoger sólo una. De esta forma, los estudiantes saben exactamente qué es lo que pueden decir para una pregunta dada. Por otra parte, los sistemas de respuesta abierta simplemente formulan una pregunta al usuario, y éste a su vez debe generar la secuencia de palabras más apropiada.

La tecnología ASR detrás de un sistema de respuesta cerrada es comparativamente más simple, ya que en cada interacción la perplejidad es bastante baja y el vocabulario es muy pequeño. Utilizando un método de reconocimiento que posea una exactitud cercana a 90 % estos sistemas tienden a ser muy robustos. Por otra parte, los sistemas de respuesta abierta son mucho más complejos y exigentes en cuanto a requerimientos técnicos y pedagógicos: basta pensar en las posibles respuestas que se pueden generar a partir de: *¿Cómo llego al aeropuerto?*. Claramente, es necesario complementar la labor del ASR con procesamiento de lenguaje natural y estrategias para evitar interpretaciones equívocas.

2.2.4. Evaluación automática de pronunciación y prosodia

La tecnología ASR ha sido aplicada también al problema de enseñanza de pronunciación, y en la actualidad es materia de investigación de muchos científicos. Un sistema de evaluación de pronunciación consiste en tutor virtual que invita a los estudiantes a repetir determinadas palabras y frases cortas con el propósito de practicar y mejorar la calidad de su lenguaje hablado, específicamente lo que respecta a la producción de sonidos (características segmentales). Para esto, se utilizan modelos acústicos que representan la pronunciación de los hablantes nativos, con los que se entrenan sistemas ASR para reconocer pronunciaciones correctas e incorrectas (Gu L. & Harris J., 2003; Neumeyer et al., 1999).

Otros sistemas se han concentrado en la evaluación de características prosódicas, principalmente la entonación (Kim Ch. & Sung W., 2004) y la acentuación o *stress* (Tepperman & Narayanan, 2005). En estos trabajos, se han destacado la importancia de los elementos suprasegmentales en el contexto de CALL, como por ejemplo la información extralingüística contendia en ellos. Entrenar la percepción y producción de la prosodia sin duda provoca un efecto positivo en estudiantes avanzados, ya que les permite enriquecer su discurso agregando fluidez y naturalidad. Por otro lado, los alumnos principiantes también podrían resultar beneficiados en la aprehensión del lenguaje, ayudándolos a superar el temor a hablar en otro idioma, y a evitar las entonaciones planas causadas por la timidez.

Indudablemente, la retroalimentación o *feedback* es clave en los sistemas de evaluación de pronunciación, debido a que el estudiante necesita una respuesta que refleje objetivamente la calidad del habla del estudiante más alla de su propia percepción.

El *feedback* consiste en un estímulo audiovisual, que depende de la habilidad que se está evaluando. Para evaluación de pronunciación, es común utilizar un puntaje o *score* cuyo valor refleja el nivel de exactitud alcanzado. Los investigadores han centrado sus esfuerzos en diseñar e implementar métodos cuyo criterio de evaluación sea lo más parecido a un profesor real (Franco H. et al., 1997). Para la evaluación de características prosódicas como la entonación, además de una nota se puede utilizar un gráfico que muestre la tendencia de la frecuencia fundamental en el tiempo, como el sistema SLIM (Delmonte R. et al., 1997). En resumen, el *feedback* es el elemento que caracteriza a CALL como un sistema de lazo cerrado.

2.3. Características suprasegmentales

Las características prosódicas de la señal de voz son también denominadas características *suprasegmentales*. Esto significa que no están confinadas a un segmento o *frame* en particular, sino que éstas ocurren a un nivel más elevado. La prosodia intenta analizar y formalizar elementos como la entonación; acentuación; duración y ritmo. Sin embargo, estos útimos no son los únicos elementos prosódicos presentes en la expresión oral, ya que pueden ser considerado como prosodia otros elementos como aspiración, nasalidad o articulación. La producción de la prosodia se asocia a los *parámetros prosódicos físicos* o correlatos acústicos constituidos fudamentalmente por el *pitch*; la energía y la duración a nivel de segmento. A continuación, se describen los elementos prosódicos más importantes en este trabajo.

2.3.1. Entonación

La entonación está definida como la combinación de las caracterísicas tonales dentro de unidades estructurales asociadas con el parámetro acústico f_0 o *frecuencia fundamental*, y sus variaciones distintivas a lo largo de una elocución (Botinis A. et al., 2001). La producción de la entonación está definida por el número de veces por segundo que las cuerdas vocales completan un ciclo de vibración.

La ecuación 2.1 muestra una relación matemática entre el *pitch* p y la frecuencia f en la escala armónica (Marchand S, 2001):

$$p(f) = p_{ref} + O \log_2 \left(\frac{f}{f_{ref}} \right) \tag{2.1}$$

En esta expresión, p_{ref} y f_{ref} corresponden a la frecuencia de un tono de referencia, mientras que O es la división de la octava. De la ecuación, es directo deducir que la diferencia entre dos armónicos consecutivos es exactamente O. Es muy usual asignar el valor $O = 12$, y de esta forma el *pitch* queda expresado en *semitonos*[1].

2.3.1.1. Representación de la Entonación

En *fonética* es común utilizar dos líneas horizontales paralelas para representar el *pitch* máximo y mínimo. Los cambios de entonación se muestran mediante líneas oblicuas o flechas, como se muestra en la Figura 2.1. Por otra parte, se utiliza una línea vertical para indicar prominencia sin cambios de entonación.

[1]En música la diferencia entre una determinada nota y su *octava* superior es igual a 12 semitonos.

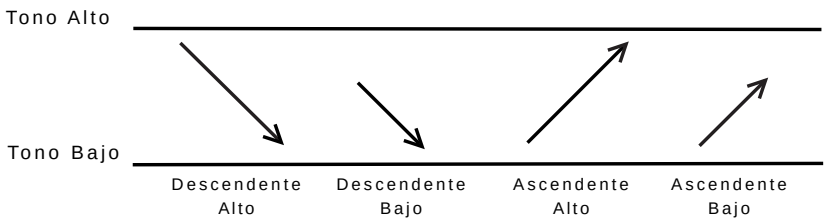

Figura 2.1: *Los principales cuatro tipos de entonación.*

2.3.1.2. Funciones de la Entonación

Entre las funciones lingüísticas de la entonación, es posible mencionar al menos seis: función actitudinal; función de prominencia; función gramatical; función de discurso; función léxico-semántica y función de naturalidad.

La Función Actitudinal hace posible expresar emociones y actitudes del habla, y además añade un tipo especial de significado al lenguaje hablado. Como muestra el siguiente ejemplo, la misma oración puede mostrar distintas actitudes, dependiendo de la entonación con la cual es pronunciada, como los siguientes saludos en inglés:

`Good ↘morning` (Saludo normal)
`Good ↗morning` (Saludo rutinario)
`Good ^morning` (Saludo cordial)
`|Good ^morning` (Saludo entusiasta)

Otro ejemplo, son las siguienes expresiones de orden:

`|Leave him a↘lone` (Orden)
`|Leave him a^lone` (Orden impaciente)

La distinción entre órdenes y peticiones recae principalmente en la entonación:

ˈCome ↘closer	(Orden)
ˈCome ˅closer	(Petición)
↘Come ↗closer	(Petición con algún grado de impaciencia)

Es posible relacionar la actitud (Austin, 1962) con la intencionalidad del hablante. De acuerdo con este autor, el significado literal tiene una cierta expresión que no siempre corresponde a lo que el hablante quiere comunicar. De esta forma, las tres frases en inglés mostradas como ejemplo anteriormente ('Good Morning'; 'Leave him alone'y 'Come closer') adquieren una distinta intención cuando son pronunciadas con distintos patrones melódicos.

La entonación tiene un rol significativo cuando se asigna *prominencia* a las sílabas que deben ser reconocidas como acentuadas. Esta sílaba, que habitualmente es pronunciada con un tono mayor, se presenta con mayor intensidad (energía). Varios autores se han referido a este tipo de acento, "Cuando se desea enfatizar (para contrastar) alguna parte en especial de una palabra que normalmente no está acentuada, dicha parte puede recibir un énfasis fuerte, y el acento primario puede llegar a convertirse en secundario" (Jones J., 1962). Otros lo han denominado "Acento de Insistencia"; "Acento enfático" (Brosnahan L. & Malmberg B., 1970) y "Acento Expresivo" (Chomisky N. & Halle M., 1968), siendo el último el más ampliamente aceptado.

Este tipo de acento es comunmente empleado para indicar contraste:

These agreements must be evaluated ↘**ob**jectively,
not ↘**sub**jectively.

También es posible utilizar este acento para indicar énfasis:

I saw an extra↘ordinary Italian film.

Cuando se usa la información dada por la entonación, es más fácil para el oyente reconocer la estructura gramatical y sintáctica de la frase u oración dicha por algún individuo. Por ejemplo, para determinar la colocación de una frase, cláusula o frase frontera, o la distinción entre frases afirmativas o interrogativas. Esta función se denomina Función Gramatical. El siguiente ejemplo en inglés ilustra esta función:

I have ↓planes to ↘leave. (Tengo planes de irme.)

I have ↘planes to leave. (Hay planes que debo dejar.)

Otro ejemplo, en español:

No ↗tengo que leer. (No tengo obligación de leer.)

No tengo ↗qué leer. (No tengo material para leer.)

Considerando que el acto de hablar desde una perspectiva amplia, es posible notar que la entonación puede sugerir al oyente qué puede ser tomado como *nueva* información, y qué es considerado como información *dada*. también puede sugerir que el hablante está indicando un tipo de contraste o enlace con algún material presente en otra unidad tonal y en una conversación, puede proveer una indicación en relación al tipo de respuesta esperada. Se muestran algunos ejemplos de la Función de Discurso de la entonación:

I sent the ↘book to John on Tuesday.

I sent the book to ↘John on Tuesday.

I sent the book to John on ↘Tuesday.

Como se puede observar, la primera frase hace énfasis en *book*, lo cual indica que se está recalcando el objeto, y no la persona a quien se envió el libro, o el día en que fue realizada la acción, como muestra la segunda y tercera frase respectivamente.

Hay algunos idiomas llamados *lenguas tonales*, donde la entonación es de vital importancia, ya que la melodía puede cambiar el significado de una palabra completamente. Este es el caso de numerosos idiomas orientales, como el Vietnamita o el Chino Mandarín. Respecto a este último, es posible citar el siguiente ejemplo que muestra como cambia el significado de la secuencia de sonidos /*ma*/, entonada de distinta forma:

˩ma	mamá
˧˥ma	cáñamo
˨˩ma	caballo
˥˩ma	regañar

En estos casos, se puede decir que la entonación cumple una Función Léxico-Semántica.

Es posible incluir una sexta función, más difícil de definir y describir, pero que es reconocible por cualquier hablante nativo competente de un determinado idioma. Una adecuada entonación provee de Naturalidad al habla. Los hablantes nativos pueden reconocer si una frase ha sido pronunciada por otro nativo o no. Hay varias características que contribuyen a esto, siendo algunas más fáciles de distinguir que otras: Elección de las palabras; estructura sintáctica; características segmentales; y ciertamente, la entonación y el ritmo. Sin embargo, un hablante extranjero competente puede desviar la atención del interlocutor, si su entonación (en conjunto con el ritmo) no es el mismo usado por un hablante nativo en las mismas circusntancias, ya que su habla se oiría poco natural.

2.3.1.3. Importancia de la entonación

La entonación es muy importante en el proceso de comunicación. Los hablantes de cualquier idioma reconocen tal importancia al hacer comentarios como: "Está de acuerdo, pero lo dijo de una manera..." . En muchas ocasiones, la "forma" en que se dice algo es más importante que el mensaje literal, la organización sintáctica o las palabras usadas. Frecuentemente, la prosodia sugiere exactamente lo contrario a las palabras usadas por el hablante. Ejemplos de este fenómeno son las oraciones sarcásticas e irónicas.

En pragmática, se habla de tres fuerzas del lenguaje (Hallyday M., 1994): locutoria; ilocutoria; y perlocutoria. La primera de ellas hace referencia a lo literalmente expresado, al significado que tiene un conjunto de palabras en una frase. La fuerza ilocutoria se refiere a la intencionalidad con la que una proposición es expresada. Finalmente, la fuerza perlocutoria es la que produce una respuesta en el receptor. Por ejemplo, la oración en español: *¿Puedes pasarme la sal?* es aparentemente una pregunta acerca de la capacidad del interlocutor para hacer algo. Esta es la fuerza locutoria o el significado literal de la expresión. No obstante, cuando esta frase es pronunciada, la intención del hablante es que alguien le acerque un salero. Esta es la fuerza ilocutoria.

Como se pudo apreciar en el ejemplo anterior, existen diferentes niveles de significado, y cualquier hablante competente reconoce estos niveles y actúa de forma apropiada. Por ejemplo, si la respuesta a la pregunta anterior fuera: "Si, evidentemente", el locutor se hubiese sorprendido. En otras palabras, no se espera una respuesta al significado literal del significado de la expresión, sino más bien una acción conforme al significado intencional. La fuerza perlocutoria correspondería por tanto a la reacción del interlocutor en respuesta de la ilocutoria. En el proceso ejemplificado anteriormente la prosodia juega un rol fundamental, particularmente en la fuerza ilocutoria.

La entonación es tan importante que incluso puede ser usada sin una palabra. El sonido /*m*/ puede ser emitido con diferentes tonos como respuesta a un comentario, por ejemplo. Puede indicar acuerdo, desacuerdo, duda, placer dependiendo de la variación tonal con el cual es pronunciado. La presencia de la prosodia en el habla delata una *deficiencia* en el lenguaje escrito. Por esta razón, es común el uso de textos en negrita, cursiva, mayúsculas y comillas para describir intenciones y actitudes.

En resumen, la prosodia es fundamental en la enseñanza y en la oratoria. Si se comparan dos hablantes que poseen un nivel similar de conocimiento de un determinado tema, y que además manejan la misma cantidad de elementos léxico-semánticos y gramaticales, la audiencia favorecerá a aquel que posea un mejor dominio de los elementos suprasegmentales. El uso correcto de pausas, de énfasis y entonación ciertamente facilitan la cohesión y coherencia del discurso.

Existen algunos autores que entienden la entonación de un determinado idioma como una entidad discreta, lo que ha llevado a muchos intentos de generar un modelo que describa la entonación. Sin embargo, para una misma lengua existen muchas variantes. Un hablante nativo de inglés o español puede determinar fácilmente si otro hablante nativo pertenece o no a una determinada ubicación geográfica dependiendo de las variaciones tonales sin ningún tipo de entrenamiento previo. Existen considerables diferencias en los patrones de entonación del español hablado en diferentes regiones del mundo, incluso dentro de un área pequeña (Face T., 2006).

Equiparar la entonación inglesa y española es una tarea imposible de realizar. Probablemente se puede comparar un dialecto de uno de estos idiomas con algún dialecto del otro. No obstante, se puede afirmar que en la mayoría de las lenguas no-tonales

se comportan más o menos del mismo modo en lo que respecta a entonación. Así, en muchos idiomas una melodía descendente se asocia a una afirmación o a una orden, y una melodía ascendente a una pregunta o una respuesta amable. Sin embargo, existen diferencias que pueden conducir a un malentendido, en particular las intenciones o actitudes del hablante, quien por ejemplo puede expresarse de modo insistente en vez de amable. Existe evidencia empírica que muestra diferencias significativas en la elección del tono o acento tonal entre hablantes nativos y no-nativos del inglés en contextos similares, los cuales generan problemas de comunicación (Ramírez D. & Romero J., 2005).

2.3.2. Acentuación

La acentuación depende de la intensidad con la cual un sonido es pronunciado, por ejemplo en la palabra "manzana", la segunda sílaba es pronunciada con mayor énfasis que la tercera. En inglés, la palabra "university" posee una acentuación mayor en la sílaba *ver*, el cual corresponde al *acento primario*. Por otra parte, la sílaba *u* que también está acentuada, pero en un grado menor y se dice que posee *acento secundario*, mientras que *ni*, *si* y *ty* se consideran no acentuadas.

2.3.2.1. Función Léxico-Semántica de la Acentuación

Dada una palabra, el lugar donde se encuentre su acento primario o principal puede cambiar radicalmente su significado. En inglés, la palabra *project* es un sustantivo si la primera sílaba está acentuada, mientras que es un verbo si el acento se ubica en la segunda. En español hay casos como: *Papa*, y *papá*; *hábito*, *habito*, *habitó*; *límite*, *limite*, *limité*; *célebre*, *celebre*, *celebré*; *depósito*, *deposito*, *depositó*; *ejército*, *ejercito*, *ejercitó*.

Como se puede apreciar, el acento marca diferencias que permiten distinguir entre

distintos tiempos verbales. Lo anterior, no se da en otras lenguas como el francés donde todas las palabras son acentuadas en la última sílaba, y en idiomas como el checo o finés, donde el acento siempre está presente en la primera.

2.3.3. Duración

En algunos idiomas, la duración tiene funciones léxico-semánticas muy relevantes, donde el largo de un fonema entrega un valor fonológico. En italiano, las palabras *vile* y *ville* poseen significados diferentes. En efecto, la primera significa villano, mientras que la segunda corresponde a villas. En finés, se encuentran palabras donde la duración de vocales e incluso de consonantes puede variar el significado: *tuli* (fuego), *tuuli* (viento), *tulli* (clientes); *muta* (barro), *muuta* (otro), *mutta* (pero), *muuttaa* (cambio). En estonio, se tienen los siguientes ejemplos: *lina* (hoja); *linna* (de la ciudad, con una n larga) y *linna* (a la ciudad, con una n más larga aún).

En el inglés o el español, la duración de un segmento tiene efectos significativos, pero no léxico-semánticos. Si se escucha una respuesta negativa con una n muy larga, se subentiende una actitud vacilante. Se puede afirmar que la distinción de las tres pronunciaciones del sonido fricativo *ch* en Chile depende parcialmente del largo. La pronunciación estándar tiene una duración de 0,14$[s]$, mientras que en la pronunciación marcada (usada por la gente jóven de la clase más alta) dura 0,07$[s]$ (Vivanco, 1998-1999).

En esta sección se han estudiado las características suprasegmentales de la señal de voz desde una perspectiva lingüística. A continuación se analiza la entonación y el acento desde el punto de vista de procesamiento de señales, donde se describe el tratamiento digital de la señal de voz para obtener sus características acústicas mediante el análisis en frecuencia, y sus características prosódicas mediante la estimación de la

frecuencia fundamental y la energía.

2.4. La voz humana

La voz se define como el sonido generado voluntariamente por el aparato fonador humano. Este aparato está compuesto básicamente por tres conjuntos de órganos: de respiración (pulmones, bronquios y tráquea); de fonación (cavidades glóticas, resonadores nasal y buco-faríngeo); y de articulación (paladar, lengua, dientes, labios, velo y mandíbula). En la Figura 2.2 se ilustran las partes del sistema de producción de voz.

Un sonido se genera cuando una corriente de aire emanada de los pulmones atraviesa los broquios y la tráquea, pasando luego por las *cuerdas vocales* o *pliegues vocales*, ligamentos elásticos que se encuentran dentro de las paredes de la laringe[2]. Este aire es amplificado y modificado por las cavidades naso-buco-faríngeas, para ser finalmente moldeado por los articuladores.

Los tipos de excitación del sistema de producción de voz humana son: fonación; exhalación; fricación; compresión; y vibración. Cuando los pliegues vocales están tensos y juntos, pero no completamente cerradas, el paso del aire por la laringe produce vibraciones y tales sonidos se denominan *tonales* o *sonoros*. De este modo, se genera una excitación de tipo fonación, y la frecuencia a la que oscilan las cuerdas vocales se denomina frecuencia fundamental f_0. Ejemplos de estos sonidos son todas las vocales del del idioma inglés, o consonantes como la $/m/$; $/n/$ y $/l/$. Por otra parte, si las cuerdas vocales están lo suficientemente separadas permitiendo la libre expulsión del aire sin generar vibraciones, se producen sonidos *sordos* o *áfonos*, como lo es la consonante $/j/$ del español.

[2]El espacio que existe entre las cuerdas vocales se denomina *glotis*.

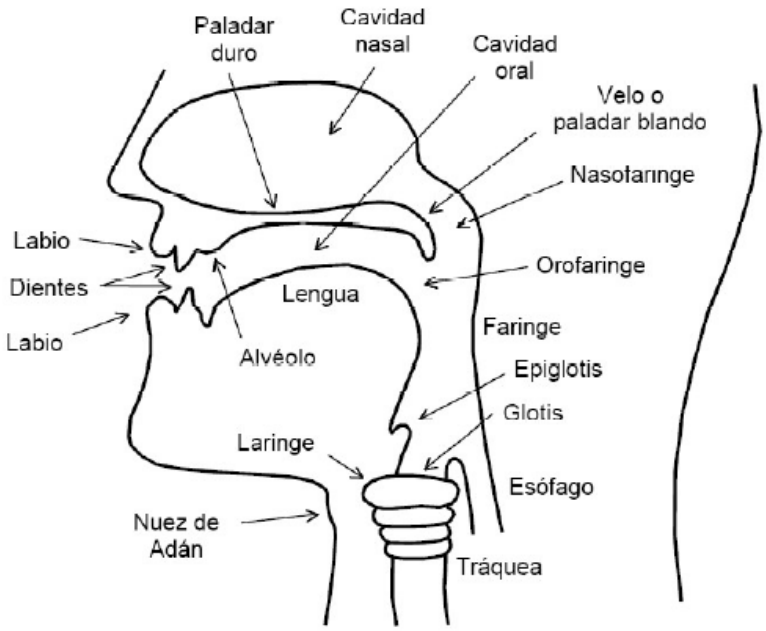

Figura 2.2: *Esquema del aparato fonador humano.*

La fricación se produce cuando el aire emitido escapa con cierta fricción producida por los órganos articulatorios como la boca o los dientes. Este tipo de excitaciones da origen a sonidos como /s/ o /f/ en español. La compresión es un tipo de excitación que se produce cuando el tracto vocal acumula presión, y luego libera el aire bruscamente, generando un ruido corto y explosivo, por ejemplo el sonido /p/. Por último, la excitación por vibración es el resultado de la fricación de un flujo de aire que simultáneamente genera vibraciones en los pliegues vocales, como sucede cuando en español se pronuncia una /b/ entre vocales.

Existen también excitaciones donde los sonidos se generan con corrientes de aire que no provienen de los pulmones. Entre ellos, se tienen los sonidos glotales, característicos del árabe y el hebreo; y los sonidos velares, comunes en las lenguas amerindias y de África oriental. Estos sonidos son poco frecuentes en las lenguas del mundo, al menos

las lenguas romances como el español no poseen ningún sonido de origen no pulmonar.

Es importante destacar que las características de la voz son dependientes de las características fisiológicas de cada individuo: el flujo de aire glotal; el tamaño de las cavidades nasales; las dimensiones y morfología del tracto vocal; la capacidad pulmonar; etc. En particular, la frecuencia fundamental que puede generar un hablante depende directamente del ancho y largo de sus cuerdas vocales, parámetros que ciertamente están relacionados con el género y la edad. En efecto, el rango de variación de f_0 para las mujeres es mayor que en los hombres adultos.

2.5. Parametrización acústica de la señal de voz

En este trabajo es de suma importancia representar las señales de voz mediante secuencias de parámetros acústicos, ya que son la entrada de los algoritmos de alineamiento que se explican más adelante. En primer lugar, se deben tener en cuenta los problemas que se presentan al modelar una elocución utilizando parámetros acústicos: variabilidad temporal; variabilidades del entorno (inter-locutor); o producidas por el mismo locutor (intra-locutor); la información extralingüística; la fuerte dependencia del micrófono; y la presencia de ruido. Se debe considerar también que la señal de voz se caracteriza por ser un proceso estocástico no-estacionario.

La conversión análogo-digital es la primera etapa en el pre-procesamiento, ya que los computadores requieren una representación discreta de la señal. En aplicaciones con interfaces hombre-máquina, esta conversión es efectuada por una tarjeta de sonido. Luego, se aplica un detector de inicio y fin de señal útil, el cual se encarga de eliminar los cuadros iniciales y finales de silencio ya que éstos no contienen información

acústica relevante. Posteriormente, se aplica un proceso de enventanado que conisiste en dividir la secuencia que representa la voz en unidades llamadas cuadros o *frames*. Para ello, se utiliza la técnica de enventanado de *Hamming* (Picone, 1993) con el fin de evitar discontinuidades al principio y al final de cada segmento.

La etapa siguiente es el análisis espectral en cada cuadro, para lo cual se utiliza la tranformada rápida de *Fourier* (FFT, *Fast Fourier Transform*). Por otra parte, el oído humano no es capaz de percibir frecuencias puntuales sino que distingue intervalos, y por esta razón se utilizan bancos de filtros. Además, la percepción acústica humana presenta un comportamiento no lineal en frecuencia, por lo tanto se hace necesario usar una escala adecuada que concentre los filtros donde la capacidad de discriminación del oído sea mayor. Para esto, se utiliza la escala de *Mel* descrita por la ecuación 2.2, para una determinada frecuencia f medida en Hertz.

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.2}$$

Los filtros corresponden a funciones triangulares, con ganancia unitaria en la frecuencia central, con un traslape de 50 % y con un ancho de banda constante en la escala de *Mel*. Después, se obtienen los coeficientes cepstrales (MFCC *Mel-cepstrum frequency coefficients*), los cuales se calculan a partir de la energía de cada filtro y la transformación discreta de coseno (DCT, *Discrete Cosine transform*). Por último, se estiman las derivadas de primer y segundo orden de los MFCC. Luego cada *frame* es acústicamente representado por un vector de coeficientes que lo identifica de manera única. La Figura 2.3 muestra un diagrama de bloques del proceso de extracción de parámetros acústicos.

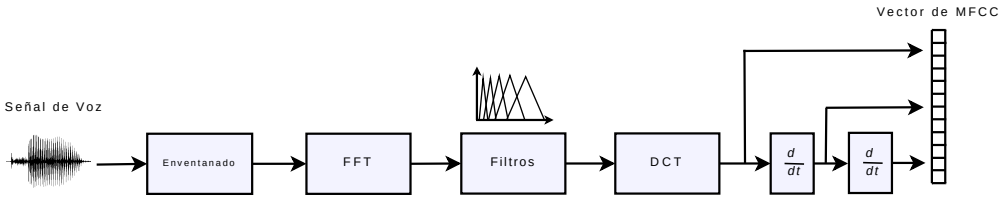

Figura 2.3: *Diagrama de bloques de la parametrización acústica de la señal de voz.*

2.6. La frecuencia fundamental f_0

2.6.1. Definiciones básicas

La frecuencia fundamental, tono fundamental o simplemente f_0 se define como la frecuencia más baja de la descomposición armónica de una señal cualquiera. Cuando se hace referencia específicamente a señales de voz, surgen varios problemas con la definición anterior debido a la no estacionariedad de éstas y a la presencia de segmentos que no son periódicos. En este contexto, se habla de sonidos *sonoros* cuando hay presencia de peridiocidad en un determinado fonema o segmento de voz, y de sonidos *sordos* o *áfonos* cuando no la hay. Por ejemplo, en el español los sonidos vocálicos $/a/$ y $/e/$, o el sonido nasal $/m/$ son considerados sonoros, mientras que los sonidos fricativos $/s/$ y $/f/$ son clasificados como sordos.

En la Figura 2.4 (a) se muestra un extracto de la forma de onda de voz de una mujer pronunciando el sonido $/e/$, en la cual se aprecia la peridiocidad de la señal, mientras que en (b) se muestra otra señal de la misma persona produciendo el sonido $/s/$, donde se puede observar que la onda es completamente aleatoria.

Existen al menos tres puntos de vista con el cual se puede definir f_0 para la señal de voz (Howard I., 1991):

1. Producción de la voz: Está relacionado directamente con la excitación de la

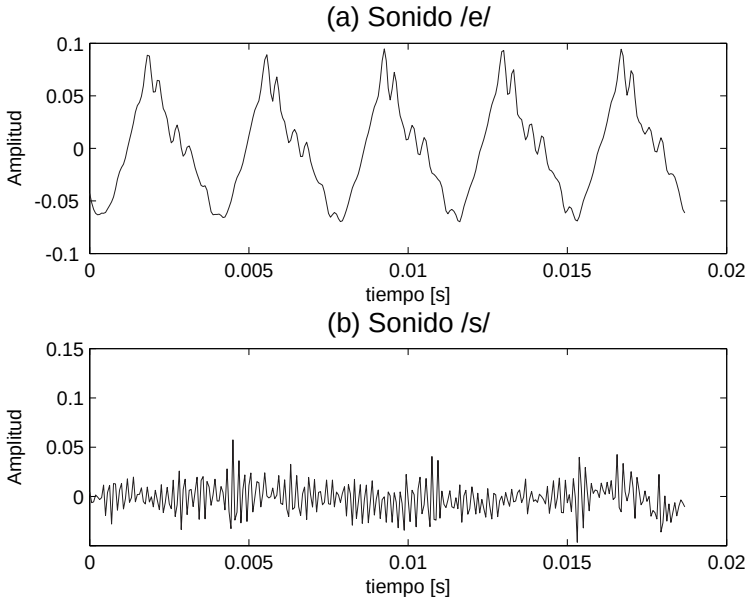

Figura 2.4: *Segmentos sonoros (a) y sordos (b) de una señal de voz femenina.*

laringe. El período fundamental t_0 se define como el tiempo transcurrido entre dos pulsos glotales sucesivos. Luego, la frecuencia fundamental está dada por:

$$f_0 = \frac{1}{t_0} \tag{2.3}$$

Por lo tanto, f_0 corresponde a la frecuencia con la cual vibran las cuerdas vocales al generar un sonido. Para captar los pulsos glotales y así obtener f_0, es necesario que el hablante se someta a una prueba llamada *laringografía*, la cual consiste en registrar los movimientos de la laringe a través de un aparato llamado laringógrafo. La Figura 2.5 muestra un ejemplo de los pulsos glotales, donde se ilustra a qué corresponde t_0.

2. Procesamiento de la señal: Se caracteriza por observar la forma de onda de

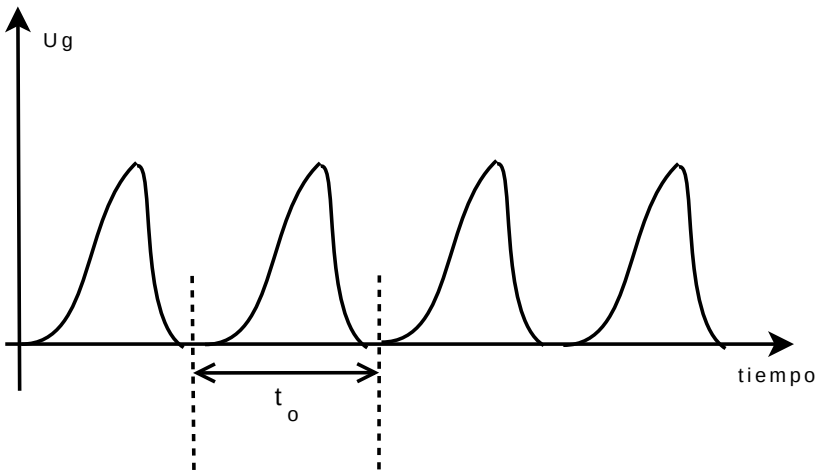

Figura 2.5: *Diagrama de pulsos glotales.*

la voz. El valor de f_0 en este caso está dado por la cantidad de ciclos cuasi-periódicos en un intervalo de tiempo determinado. Desde el punto de vista matemático, esta es la forma óptima para abordar el análisis de f_0.

3. Percepción: Está relacionada más bien con el dominio de la frecuencia, y con la forma en que el oído humando la percibe. En este caso se habla de *pitch*, el cual se define como la frecuencia de aquella sinusoide que evoca la misma altura percibida por un segmento de voz. En estricto rigor el *pitch* no corresponde a la frecuencia fundamental, sino que se entiende como un fenómeno subjetivo relacionado con la percepción de f_0, y no a un parámetro de la producción de voz. A pesar de las definiciones acústicas y perceptuales de f_0 y *pitch* respectivamente, estos términos son utilizados indistintamente en la literatura.

2.6.2. Estimación de la frecuencia fundamental f_0

La estimación de la frecuencia fundamental ha sido por muchos años un tema de investigación muy importante, y continúa siéndolo en la actualidad debido a su gran

cantidad de aplicaciones como por ejemplo síntesis de voz; codificación de voz; reconocimiento de género mediante la voz; entre otras. Las etapas básicas en la estimación del *pitch* en señales de voz, las cuales se ilustran en la Figura 2.6.

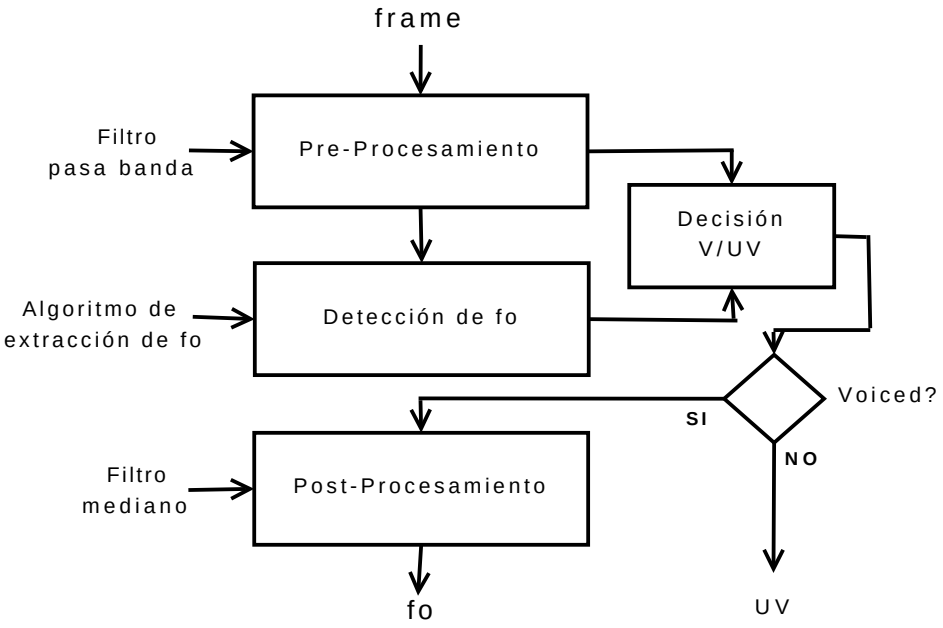

Figura 2.6: *Etapas en detección de f_0.*

2.6.2.1. Dificultades asociadas a la detección de pitch

La detección de f_0 en una señal de voz es al parecer una tarea relativamente simple, pero existe un gran número de dificultades asociadas. A continuacion se explican algunas de ellas.

- Es muy difícil evaluar la calidad de un sistema o método de detección de f_0, ya que no se tiene una medida objetiva de f_0 para una señal cualquiera.
- Dada una elocución, es necesario saber cuáles son los *frames* que contienen trozos

sonoros y sordos, ya que en estos últimos no tiene sentido calcular f_0 y se considera que no hay *pitch*. Esto se conoce como *Voiced/Unvoiced Detection*.

- Los segmentos sordos y sonoros a lo largo de la señal se presentan de forma bastante irregular, por tanto se puede incurrir en errores de este estilo que convergen en inserciones erroneas y omisiones de f_0 en segmentos sonoros.
- El tracto vocal o el canal de transmisión pueden enfatizar otros armónicos, provocando que alguno de ellos sea detectado como f_0.
- Es muy común que armónicos o sub-armónicos sean detectados como frecuencia fundamental, fenómeno que se conoce como *saltos de octava*.
- El ruido que proviene de la respiración del hablante hace que los segmentos sonoros parezcan áfonos.
- Cuando se utilizan filtros de banda muy angosta, los segmentos sordos pueden presentar una peridiocidad aparente.

2.6.2.2. Pre-procesamiento

En esta etapa se intenta eliminar elementos que interfieren en la señal como el ruido; influencia del tracto vocal y del canal de transmisión; presencia de componentes continuas; etc. El uso de filtros introduce mejoras importantes en la detección de f_0, ya que la banda de frecuencias de interés está bien determinada $(80 - 600[Hz])$. De esta forma se eliminan aquellas componentes que se encuentren fuera de tal rango, por ejemplo el ruido de baja frecuencia de las fuentes de poder, el nivel DC (componente de frecuencia igual a cero) y el ruido de soplido de micrófono.

Antes de proceder a estimar el *pitch* con cualquier método, es necesario clasificar a los *frames* como sordos o sonoros, para así reducir los costos de procesamiento y la

posibilidad de detectar f_0 erróneos. Existen trabajos que se dedican exclusivamente a resolver este problema (Kobatake H., 1987), que también se conoce como decisión V/UV (*Voiced/Unvoiced*). Por otra parte, es posible que la decisión V/UV esté implícita en una cierta técnica de estimación de f_0 (Alkulaibi A. et al., 1996).

2.6.2.3. Métodos de estimación de f_0 en el dominio del tiempo

Existe una familia relativa al dominio del tiempo para la detección de f_0 que intenta contar la cantidad de eventos que se repiten en una señal cuasi-periódica. Los métodos más importantes se explican a continuación:

- Tasa de cruces por cero (ZCR, *Zero Crossing Rate*), la cual consiste en medir la cantidad de veces que la señal pasa por cero por unidad de tiempo. El análisis espectral usando ZCR fue propuesto por (Kedem B., 1986), y fue uno de los primeros métodos usados para calcular la frecuencia fundamental, dando buenos resultados para señales que tienen concentrada la densidad espectral de potencia en f_0, ya que una onda en estas condiciones pasa dos veces por cero en un período. Sin embargo, la presencia de armónicos y componentes de otras frecuencias hacen que en un ciclo existan varios cruces, por lo tanto esta metodología presenta varias desventajas cuando se analizan señales de voz en condiciones adversas.

- Autocorrelación (Rabiner L., 1977). La correlación entre dos formas de onda es una medida de su simlitud. Entonces, si se calcula la correlación de esta onda sobre sí misma con distintos desplazamiento temporales, se obtienen valores que representan la similitud para cada uno de ellos. Eso se define como *autocorrelación*, y su definición matemática se muestra en la siguiente expresión:

$$R_x(\nu) = \sum_{n=-\infty}^{\infty} x(n)x(n+\nu) \tag{2.4}$$

Aquí, x_n corresponde a una función discreta, y ν es el desplazamiento temporal. La función $R_x(\nu)$ tendrá máximos locales cuando $x(n)$ es similar a $x(n+\nu)$, en efecto, si $x(n)$ tiene período P entonces tales máximos están en $\nu = lP$, siendo l un entero. Intuitivamente, el primer máximo está en $R_x(0)$, y el siguiente en $R_x(P)$. Así, el período fudamental está dado por el valor de ν que genera el segundo máximo local. En la Figura 2.7 se muestra la función de autocorrelación para el segmento de voz mostrado en la Figura 2.4 (a).

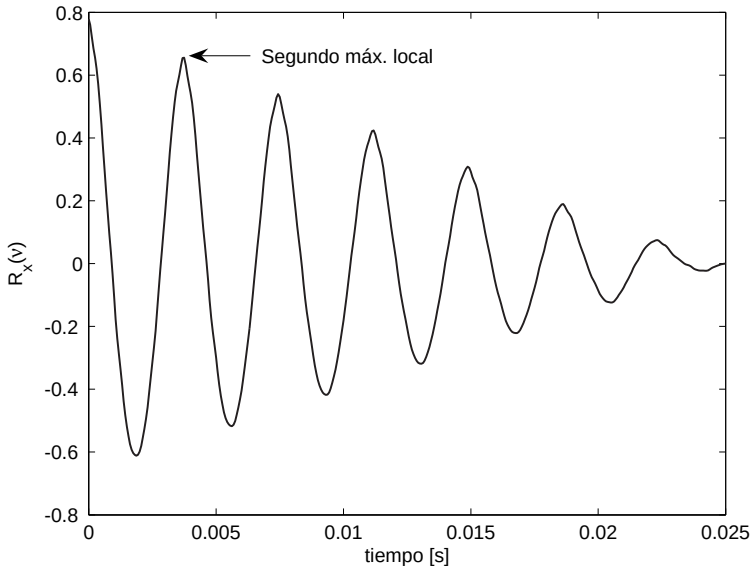

Figura 2.7: *Función de autocorrelación para un frame.*

Este método ha sido considerado por muchos investigadores debido a su simplicidad y buen comportamiento frente a ambientes ruidosos (Oh K. & Un C., 1984). Existen muchas variantes, entre ellas la utilización de una función ponderada para calcular la autocorrelación (Shimamura T. & Kobayashi H., 2001; Hung W., 2002); la aplicación de métodos de codificación de voz (Burnett I. & Gambino

P., 1996), o la elección de tamaños de frames consecuentes con el desplazamiento temporal (Hirose K. et al, 1992).

Sin embargo, esta familia de algoritmos presenta ciertos problemas debido al tamaño de la ventana de análsis y a la presencia de falsos segundos máximos en la autocorrelación debido a la presencia de componentes de alta frecuencia (Rabiner L., 1977).

- El estimador YIN. Fue introducido por (de Cheveigné A. & Kawahara H., 2002), e inspirado en el método anterior. La idea es utilizar una función similar a la autocorrelación, tratando de encontrar aquel desplazamiento temporal que minimiza la diferencia entre $x(n)$ y $x(n+\nu)$, en vez de maximizar su producto:

$$d_t(\nu) = \sum_{n=1}^{W} (x(n) - x(n-\nu))^2 \tag{2.5}$$

Para reducir la presencia de errores producidos por subarmónicos (por ejemplo, cuando la potencia del primer formante es muy alta), YIN utiliza una función acumulada normalizada, que atenúa estos efectos indeseados:

$$d'_t(\nu) = \begin{cases} 1 & \text{si } \nu = 0 \\ \frac{d_t(\nu)}{\frac{1}{\nu}\sum_{j=1}^{\nu} d_t(j)} & \text{en otro caso} \end{cases} \tag{2.6}$$

Finalmente, se añaden etapas de interpolación y de eliminación de errores de octava. El lector puede obtener mayores detalles en el artículo citado.

2.6.2.4. Métodos de estimación de f_0 en el dominio de la frecuencia

La mayoría de los métodos que están en este grupo usan la transformada rápida de Fourier (FFT, *Fast Fourier Transform*) y trabajan sobre el espectro de la señal de voz.

- Análisis de Cepstrum. El cepstrum corresponde a la transformada inversa de Fourier del logaritmo del espectro de una señal. Si la amplitud del logaritmo del espectro contiene armónicos regularmente espaciados, entonces el análisis de Fourier del espectro mostrará un *peak* correspondiente al espacio entre armónicos, es decir, la frecuencia fundamental. En resumen, la idea es encontrar peridiocidad en el espectro de la señal, a través de máximos locales en el dominio cepstral. En la Figura 2.8 se muestran los pasos a seguir para encontrar f_0 mediante el análisis descrito.

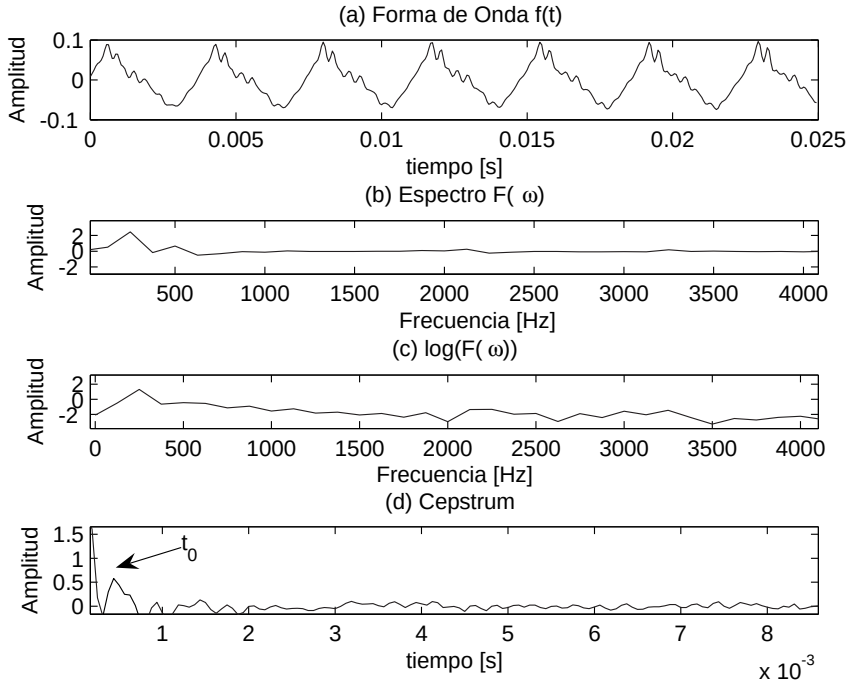

Figura 2.8: *Etapas del análisis de cepstrum.*

Este método asume que los armónicos de la señal están regularmente espaciados

en el dominio de la frecuencia. Si este no fuera el caso, tal como ocurre cuando se tiene una sinusoide pura u otra señal sin armónicos, el presente método arrojaría resultados erróneos. Por lo tanto, el análisis de cepstrum está catalogado como un procedimiento que puede aplicarse sólo a ciertos tipos de señales.

- MULTIRESOLUCIÓN. Corresponde a una mejora que puede ser aplicada a cualquier estimador de la frecuencia fundamental que trabaja en el dominio espectral. Si se tiene una estimación de f_0 de un determinado algoritmo, es posible aceptar o rechazar dicha hipótesis usando el mismo algoritmo, pero a una resolución distinta, esto es, usando una ventana más grande o pequeña para efectuar el análisis de Fourier. Si se tiene un mismo valor para distintas resoluciones, esto es una confrmación de la validez de la estimación. Este procedimiento suele arrojar resultados muy buenos, pero es muy costoso en términos de procesamiento computacional.

- PRODUCTO DE ARMÓNICOS EN EL ESPECTRO. Esta técnica, que también es conocida como HPS (*Harmonic Product Spectrum*), consiste en medir la máxima coincidencia de los armónicos de acuerdo con la ecuación 2.7 para cada *frame*.

$$Y(\omega) = \prod_{r=1}^{R} |X(\omega r)| \tag{2.7}$$

En esta ecuación, R es el número de armónicos a considerar, y ω_i es el rango de las posibles frecuencias fundamentales. Luego, al arreglo resultante $Y(\omega)$ se le calcula el máximo $\hat{Y}$ (Ecuación 2.8). Así, del argumento que maximiza la expresión se deduce la frecuencia fundamental.

$$\hat{Y} = \max_{\omega_i} \{Y(\omega_i)\} \tag{2.8}$$

Las ventajas de este método son: bajo costo computacional; comportamiento razonable frente a ruido aditivo y multiplicativo; y ajustable a distintos tipos de entrada. Por ejemplo, se puede crear una variante del algoritmo por medio de una suma de espectros, en vez de efectuar un producto de éstos.

- Máxima Verosimilitud (MacAulay R., 1978). Este algoritmo busca entre un conjunto de posibles $\tilde{X_\omega}$ y escoge aquel que mejor calza con la forma del espectro de entrada X. Tales $\tilde{X_\omega}$ están definidos como un tren de impulsos de cierta frecuencia ω, la cual es convolucionada con la transformada de Fourier del *frame* en cuestión.

  Sea $E(\omega)$ el error que se comete al aproximar $X(\omega)$ por el espectro ideal:

$$\begin{aligned} E(\omega) &= \left\| X - \tilde{X_\omega} \right\|^2 \\ \Rightarrow E(\omega) &= ||X||^2 + ||\tilde{X_\omega}||^2 - 2X{\tilde{X_\omega}}^T \end{aligned} \tag{2.9}$$

  Como los términos $\|X\|^2$ y $||\tilde{X_\omega}||^2$ son constantes, para encontrar el estimador más verosímil de $\hat{X}$ se debe minimizar el error cuadrático:

$$\hat{X} = \min_{\omega} \{E(\omega)\} = \max_{\omega} \left\{ X\tilde{X}_\omega^T \right\} \tag{2.10}$$

2.6.2.5. Post-procesamiento

Una vez que ya se ha estimado el *pitch* con algún método de los revisados anteriormente, es posible aplicar una última etapa con el fin de eliminar algunos errores producto del ruido o corregir problemas propios de cada técnica. Cabe aclarar que la estimación de f_0 se realiza *frame* a *frame*, mientras que el post-procesamiento (también denominado *tracking*) trabaja sobre todos los valores de f_0 estimados, entendiendo a la frecuencia fundamental como función del tiempo.

La Figura 2.9 muestra un ejemplo de los posibles problemas que ocurren después de haber calculado f_0 para cada *frame*. En el gráfico superior se muestra la frecuencia fundamental de referencia en función del tiempo para una determinada señal de voz, mientras que el gráfico inferior contiene los f_0 calculados con un cierto algoritmo. Se pueden apreciar cuatro tipos de problemas: detección de f_0 en segmentos sordos (a); saltos a octava superior o *doubling* (b) es decir, cuando se detecta el primer armónico como frecuencia fundamental; omisión en segmentos sonoros (c); y saltos a octava inferior o *halving* (d), esto es, la detección del subarmónico de f_0.

El uso de filtros medianos ha sido un método tradicional de post-procesamiento (Rabiner L. et al., 1975), el cual elimina f_0 erróneos y además produce un efecto suavizado o *smoothing*. Además, existen técnicas más robustas como aplicación de métodos estadísticos (Yong Duk Cho K. et al., 2002); heurísticos (Zaho X. et al., 2007) y de análisis en el dominio de la frecuencia (Marchand S., 2001).

2.7. La Energía

La energía de un *frame* está dada por la siguiente expresión:

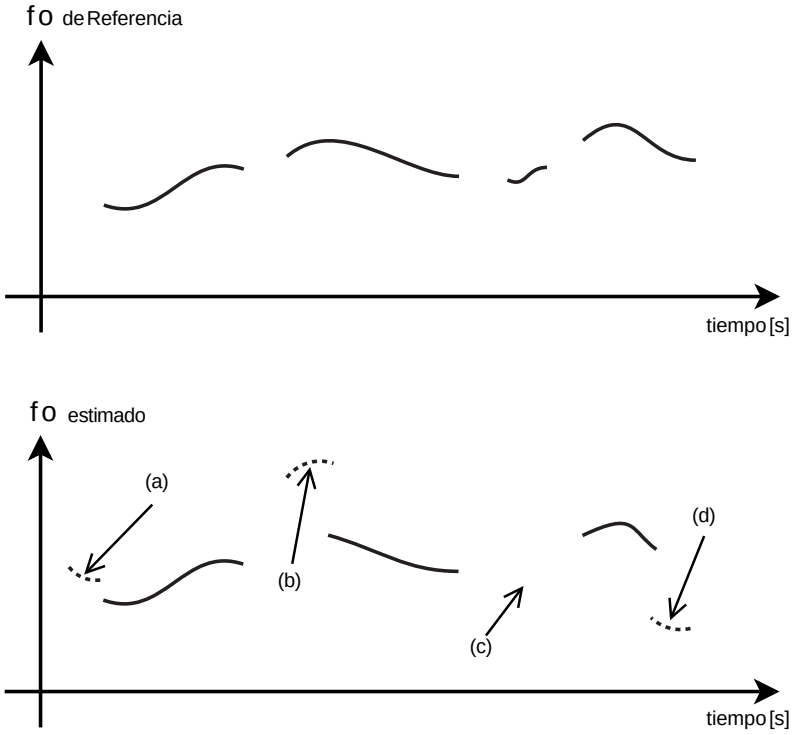

Figura 2.9: *Errores en la estimación de* f_0.

$$E_f = \sum_{n=1}^{N_m} s_n^2 \tag{2.11}$$

Esta característica contiene información importante acerca del contenido fonético, ya que entrega una idea de la intensidad con la que el hablante emite un determinado segmento de voz. En el procesamiento de la señal de voz, el uso de la energía es muy útil, y sus aplicaciones incluyen detección de inicio y fin de una señal; decisión sordo/sonoro; y otras. Para la presente investigación, lo más relevante es que la energía está fuertemente correlacionada con aspectos del habla como la acentuación.

2.8. El algoritmo DTW

Cuando se desea tener una idea de similitud entre un objeto de referencia X y uno de prueba Y, lo más natural es definir una distancia y a través de ésta cuantificar el grado de semejanza entre ellos. El problema surge cuando tales objetos están representados por distinta cantidad de vectores, ya que no es posible aplicar medidas de distancia convencionales como la euclidiana o Mahalanobis.

La situación descrita anteriormente se conoce como el problema de *Alineamiento Temporal*, el cual aparece cuando se desea comparar dos señales de voz. Una palabra o frase no puede ser pronunciada siempre a una misma velocidad, ni reproducirse de igual forma por dos locutores distintos. En la Figura 2.10 se muestra en forma gráfica un intento de comparación de dos curvas morfológicamente similares pero de largo distinto, utilizando una medida de distancia clásica, y usando alineamiento temporal.

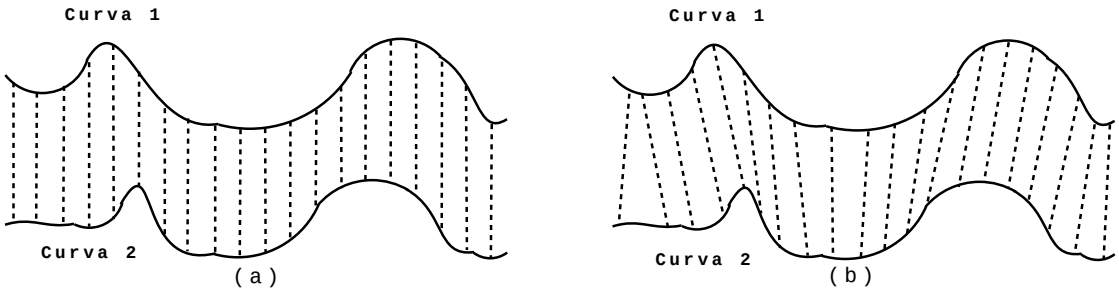

Figura 2.10: *Comparación de curvas punto a punto (a), y usando alineamiento temporal (b).*

El algoritmo de Alineamiento Temporal Dinámico (DTW, *Dynamic Time Warping*) intenta resolver el problema de comparar dos objetos que eventualmente pueden tener distinta longitud. La idea básicamente es modificar de forma no lineal la dimensión temporal de ambos objetos, y mapearlos a un único conjunto de índices, de tal forma

que la distancia calculada componente a componente sea mínima. Así, las zonas donde el objeto de referencia X se parece al objeto de prueba Y quedan "alineadas".

En la Figura 2.11 se observan dos series de tiempo, cuyos ejes de tiempo definen una grilla sobre la cual se pueden dibujar todos los posibles alternativas de alineamiento entre ambas. De color oscuro se encuentra demarcado el camino que minimiza la distancia entre las series, denominado *camino óptimo*.

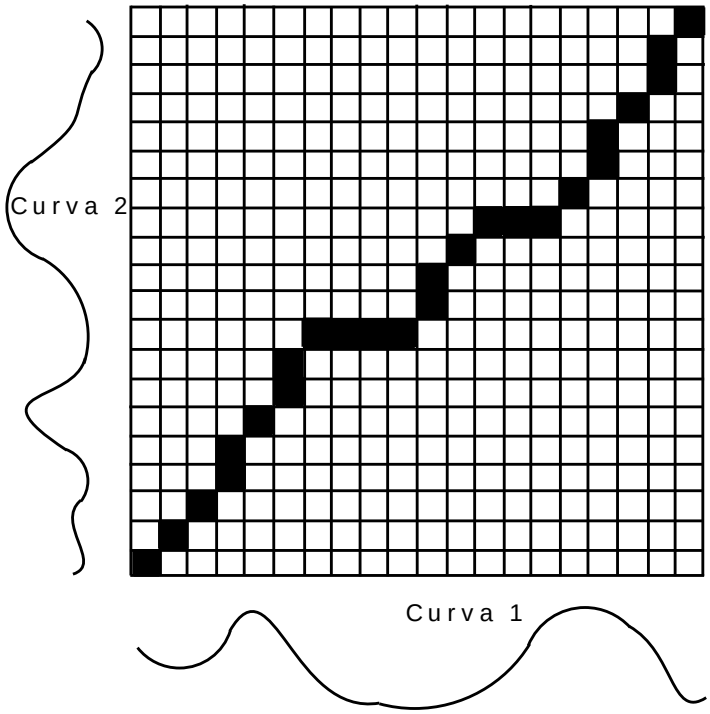

Figura 2.11: *Distancia normalizada en el tiempo entre dos curvas.*

2.8.1. Planteamiento del algoritmo

Sean dos señales de voz X y Y, las cuales pueden ser representados mediante secuencias $(x_1, x_2, ..., x_{T_x})$ y $(y_1, y_2, ..., y_{T_y})$, donde x_i e y_i son vectores de características (notar las secuencias no tienen el mismo largo). Por otra parte, se definen dos funciones

de alineamiento ϕ_x y ϕ_y, las cuales transforman los índices de las secuencias de vectores a un eje k normalizado en el tiempo:

$$\begin{aligned} i_x &= \phi_x(k), k = 1, 2, ..., T \\ i_y &= \phi_y(k), k = 1, 2, ..., T \end{aligned} \tag{2.12}$$

Lo anterior entrega un mapeo de $(x_1, x_2, ..., x_{T_x})$ a $(x_1, x_2, ..., x_T)$, y de $(y_1, y_2, ..., y_{T_y})$ a $(y_1, y_2, ..., y_T)$. Con este mapeo, se define una distancia $d_\phi(X, Y)$, la cual se calcula mediante la siguiente expresión:

$$d_\phi(X, Y) = \frac{\sum_{k=1}^{T} d(\phi_x(k), \phi_y(k)) w(k)}{\sum_{k=1}^{T} w(k)} \tag{2.13}$$

En esta expresión, $d(x, y)$ es una medida de distancia entre dos vectores y $w(k)$ representa pesos que se asignan al cálculo de cada distancia. De acuerdo con la ecuación 2.13, el valor de la distancia d_ϕ depende exclusivamente de las funciones de alineamiento que se escojan. Luego, el camino óptimo está dado por la siguiente ecuación (sujeta a ciertas restricciones):

$$d(X, Y) = \min_{\phi} d_\phi(X, Y) \tag{2.14}$$

2.8.2. Restricciones

Como se puede observar en la Figura 2.11, el número de caminos posibles es muy grande, por lo que es necesario imponer ciertas restricciones en la búsqueda (Sakoe H. & Chiba S., 1978), las cuales se describen a continuación.

1. Monotonía, es decir $\phi_x(k) \leq \phi_x(k+1)$ y $\phi_y(k) \leq \phi_y(k+1)$, para todo k. De esta forma, se evita que el camino del alineamiento retroceda en el tiempo.

2. Continuidad, implica que $\phi_x(k+1) - \phi_x(k) \leq 1$ y $\phi_y(k+1) - \phi_y(k) \leq 1$, $\forall k$. La idea es que no existan saltos en el tiempo, y así no se omitan características importantes.

3. Condiciones de Borde. Básicamente, consisite en imponer los puntos extremos del camino: $\phi_x(1) = 1$; $\phi_y(1) = 1$; $\phi_x(T) = T_x$ y $\phi_x(T) = T_y$. Así, se garantiza que el alineamiento no se concentra en una zona en particular.

4. Ventana de Alineamiento. Matemáticamente, $|\phi_x(k) - \phi_y(k)| \leq r$ para todo k, donde r es un entero positivo, cuyo valor se conoce como *ancho* o *radio* de la ventana de Sakoe-Chiba. Existen además otro tipo de restricciones, como el paralelógramo de Itakura, como se puede ver en la Figura 2.12.

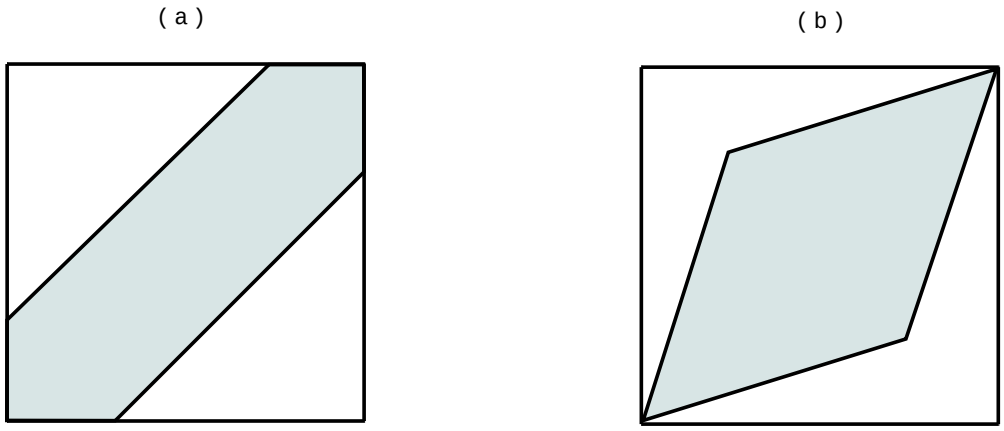

Figura 2.12: *Restricciones globales en DTW: Banda de Sakoe y Chiba (a) y paralelógramo de Itakura (b).*

5. Pendiente. Si se tiene una función de alineamiento ϕ, de tal forma que $\phi_x(k) = \phi_x(k+1)$ (o bien para $\phi_y(k) = \phi_y(k+1)$) para varios valores consecutivos de k, se traduce en que un patrón del objeto X se alinea con un segmento relativamente

grande del objeto Y. Para evitar este problema, se imponen condiciones sobre la primera derivada de ϕ, permitiendo una cantidad limitada de movimientos horizontales y verticales. En la Figura 2.13 se muestra en forma gráfica cómo se imponen las condiciones sobre la pendiente de la función de alineamiento.

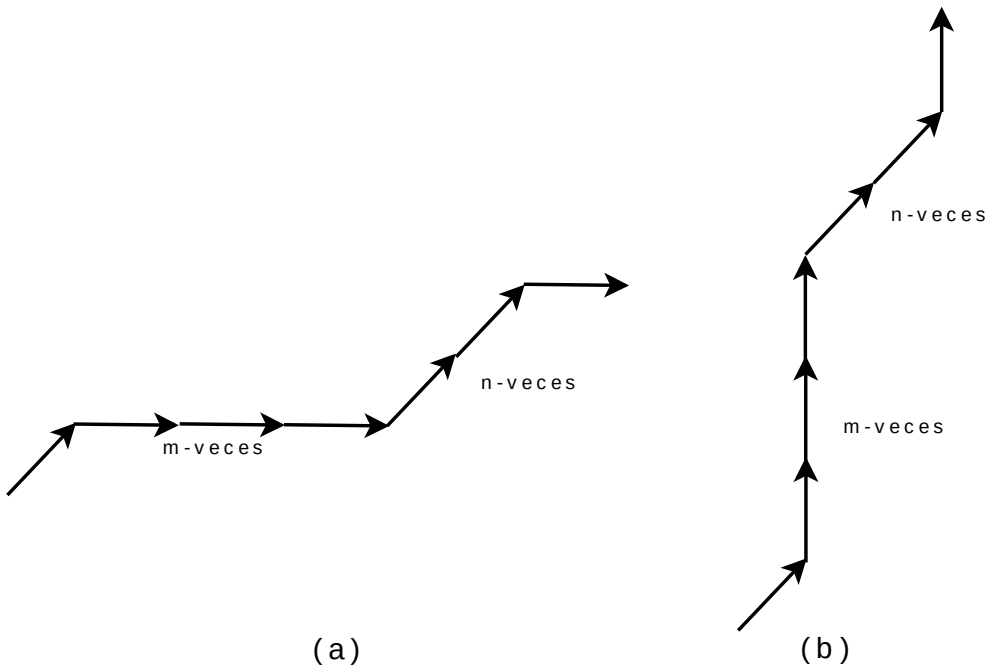

Figura 2.13: *Pendiente de la función de alineamiento: Pendiente mínima (a), y pendiente máxima (b).*

Se puede observar que (a) y (b) ilustran la pendientes mínima y máxima respectivamente. La función de alineamiento puede avanzar m veces consecutivas hacia adelante (o hacia la izquierda), luego, no puede volver a dar un paso en estas direcciones, hasta que haya avanzado n veces en la dirección diagonal.

La intensidad efectiva se representa mediante el valor $P = n/m$, que cuantifica la rigidez de la restricción. Si $P = 0$, entonces no existen restricciones en la pendien-

te de ϕ, mientras que si $P \to \infty$ el alineamiento es estrictamente diagonal. En la práctica, esta restricción se aplica sobre la ecuación de programación dinámica, que se muestra más adelante.

2.8.3. Elección de los Pesos

La ecuación 2.13 es una expresión racional, por tanto, resolver el problema de optimización que plantea la ecuación 2.14 es extremadamente complejo. Si se define W_ϕ como:

$$W_\phi = \sum_{k=1}^{T} w(k) \tag{2.15}$$

entonces, el problema planteado por la ecuación 2.14 se transforma en:

$$d(X,Y) = \frac{1}{W_\phi} \min_{\phi} \sum_{k=0}^{T} d(\phi_x(k), \phi_y(k)) w(k) \tag{2.16}$$

La ecuación anterior se puede resolver mediante técnicas de programación dinámica. Existen básicamente dos formas de elegir los coeficientes $w(k)$, de tal forma que esta simplificación sea posible.

- Forma Simétrica. Los pesos son escogidos como $w(k) = (\phi_x(k) - \phi_x(k-1) + (\phi_y(k) - \phi_y(k-1)))$. Así, $W_\phi = T_x + T_y$.
- Forma Asimétrica. Se elige $w(k) = \phi_x(k) - \phi_x(k-1)$ o equivalentemente $w(k) = \phi_y(k) - \phi_y(k-1)$. Luego, $W_\phi = T_x$, o $W_\phi = T_y$.

2.8.4. Ecuación de Programación Dinámica

El algoritmo que se describe a continuación pretende resolver la ecuación 2.16, que corresponde a un típico problema de optimización resuelto mediante programación dinámica. Sea $D(i_x, i_y)$ el *valor óptimo acumulado*, cuyo valor inicial está dado por $D(1,1) = 2d(1,1)$ para la forma simétrica. Los restantes valores se calculan mediante la recursión:

$$D(i_x, i_y) = \text{mín} \left\{ \begin{array}{l} D(i_x, i_y - 1) + d(i_x, i_y) \\ D(i_x - 1, i_y - 1) + 2d(i_x, i_y) \\ D(i_x - 1, i_y) + d(i_x, i_y) \end{array} \right\} \tag{2.17}$$

Por otra parte, si se utiliza la forma asimétrica para los pesos, entonces la condición inicial es $D(1,1) = d(1,1)$ y la recurrencia es:

$$D(i_x, i_y) = \text{mín} \left\{ \begin{array}{l} D(i_x, i_y - 1) \\ D(i_x - 1, i_y - 1) + 2d(i_x, i_y) \\ D(i_x - 1, i_y) + d(i_x, i_y) \end{array} \right\} \tag{2.18}$$

Las ecuaciones 2.17 y 2.18 no incluyen las restricciones de pendiente, luego corresponden a las de tipo $P = 0$. Si se desea imponer algún tipo de restricción de este estilo, es posible incluirla en la ecuación de programación dinámica. Por ejemplo, para imponer condiciones de tipo $P = 1$, se puede establecer para la forma simétrica una ecuación recursiva de dos pasos:

$$D(i_x, i_y) = \text{mín} \left\{ \begin{array}{l} D(i_x - 1, i_y - 2) + 2d(i_x, i_y - 1) + d(i_x, i_y) \\ D(i_x - 1, i_y - 1) + 2d(i_x, i_y) \\ D(i_x - 2, i_y - 1) + 2d(i_x - 1, i_y) + d(i_x, i_y) \end{array} \right\} \tag{2.19}$$

Usando 2.17, el alineamiento puede desplazarse libremente, hacia arriba, en diagonal o a hacia adelante tantas veces como sea necesario, mientras que mediante 2.19, el avance horizontal o vertical está condicionado por la adición de un paso diagonal. La Figura 2.14 ilustra este hecho.

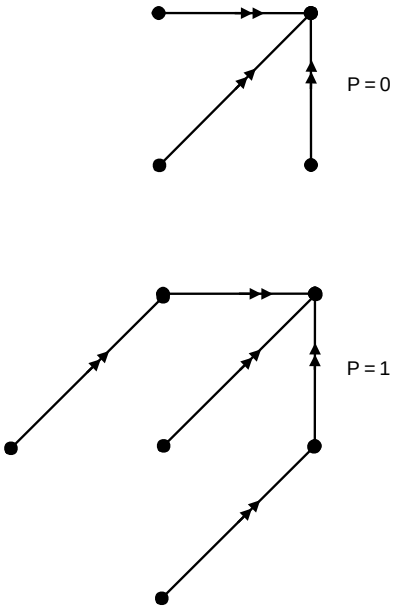

Figura 2.14: *Resricciones locales, de tipo $P = 0$ (arriba) y $P = 1$ (abajo).*

Lo anterior muestra que la restricción (5) está incluida en la ecuación de programa-

ción dinámica, mientras que la monotonía, continuidad y condiciones de borde resultan ser trivialmente cumplidas por la misma. La restricción de la ventana de alineamiento se debe ser efectuada manualmente, en la medida que se van llenando los valores de $D(i_x, i_y)$, imponiendo una distancia muy grande en los puntos que se encuentran fuera de la ventana.

2.8.5. Aplicaciones

El algoritmo DTW ha sido ampliamente usado en reconocimiento automático de voz (ASR, *Automatic Speech Recognition*), usado tanto para palabras aisladas (Sakoe H. & Chiba S., 1978; Brown M. & Lawrence R., 1982) como para habla continua (Silverman F. & Morgan D., 1990). Es considerado como el primer método que condujo a resultados aceptables en esta área. No obstante, con el pasar de los años ha sido desplazado por un nuevo paradigma llamado modelos ocultos de Markov (HMM, *Hidden Markov Models*), el cual entrega mayor robustez y escalabilidad a los sistemas ASR.

En virtud de lo anterior, el desarrollo de DTW en los últimos años no se ha centrado exclusivamente en el reconocimiento de voz sino que más bien se ha extendido su uso a otros campos de la ciencia, donde existe una gran cantidad de aplicaciones. A continuación, se muestran algunos ejemplos.

- Procesamiento de Señales: Más allá del procesamiento del habla, las aplicaciones de DTW se extienden a procesar otro tipo de señales, como por ejemplo vocalizaciones de ballenas asesinas (Brown J. et al., 2006).
- Bioinformática: Clasificación automática de cromosomas (Legrand et al., 2008); Expresión de genes (Criel J. & Tsiporkova E.).
- Minería de Datos: DTW es una herramienta ampliamente usada para comparar series de tiempo (Keogh E. & Pazzani M., 1999).

- MEDICINA: Reconocimiento de patrones en exámenes médicos, como electro-encéfalogramas (EEG) o electro-cardiogramas (ECG) (Tuzcu V. & Nas S, 2005).
- PROCESAMIENTO DE IMÁGENES: Verificación y reconocimiento de escritos a mano, como por ejemplo firmas (Shankera P. & Rajagopalan A, 2007; Faundez-Zanuy M., 2007). También se ha utilizado en el reconocimiento de movimientos y gestos de humanos (Gavrila D. & Davis L., 1995).
- QUÍMICA: Detección de patrones distorsionados, aplicados a supervisión de bio-procesos (Gollmer K. & Posten C., 1995).

2.9. Conclusiones

En este capítulo se han revisado los aspectos más relevantes del procesamiento de señales para este trabajo. Primeramente, se estudiaron los principales conceptos de la enseñanza de segundo idioma asistida por computador, donde se destacó la importancia de la prosodia. Luego, se analizaron las características suprasegmentales como la entonación, el acento y la duración desde la perspectiva lingüística. Después, se estudiaron los parámetros acústicos de la señal de voz relevantes de la prosodia, especialmente la frecuencia fundamental f_0, y por último se describió detalladamente el algoritmo de alineamiento temporal dinámico DTW.

Después de haber revisado todos los conceptos y métodos descritos aquí, el lector debiera poseer base teórica suficiente para entender el siguiente capítulo, donde se explica detalladamente el sistema propuesto. Dados estos antecedentes, el aporte de este trabajo es diseñar un sistema que utiliza varias técnicas bien conocidas del procesamiento de voz para ser aplicada a la evaluación de la prosodia en la enseñanza de segundo idioma.

Capítulo 3

El Sistema de evaluación de entonación

3.1. Introducción

El presente capítulo muestra en detalle la implementación de un método de evaluación de entonación y acentuación para CALL. La técnica se basa en la comparación de la prosodia de dos señales de voz, que corresponden a la de un usuario y una elocución pregrabada de referencia. La última parte de este capítulo muestra una serie de experimentos realizados con el fin de evaluar el desempeño del sistema para distintas configuraciones, y estudiar su comportamiento frente a distintas situaciones.

3.2. Interacción del usuario con el sistema

El sistema mostrado en este trabajo está diseñado para ayudar a estudiantes del inglés como segundo idioma a mejorar su entonación y acentuación. Esto se logra comparando la señal de voz del alumno con una elocución canónica pregrabada de un

hablante nativo, las cuales se denominan **señal de test** y **señal de referencia** respectivamente. La dinámica de interacción entre el estudiante y el sistema se explica en la Figura 3.1.

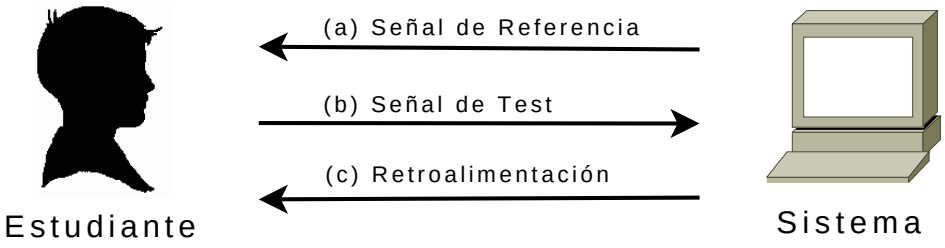

Figura 3.1: *Interacción entre el alumno y el sistema.*

El alumno escucha la señal de referencia (a), y luego responde grabando su propia voz (b), tratando de imitar o emular la entonación (o acentuación) de la señal que acaba de oir. Finalmente, el estudiante recibe algún tipo de retroalimentación o *feedback*, que puede ser de naturaleza visual (puntaje, nota, gráfica) o bien auditiva. Dado que el ejercicio anterior se puede repetir varias veces y para distintos patrones de entonación (o acentuación), el estudiante tiene la posibilidad de practicar sin la necesidad de un profesor.

3.3. Descripción del sistema propuesto

La idea central de este trabajo es diseñar e implementar un sistema que determine si dos señales de voz (referencia y test) poseen una entonación similar o no, bajo el supuesto de que ambas elocuciones pronuncian la misma frase o palabra. Para ello, se propone un sistema cuya descripción se muestra en la Figura 3.2.

En primer lugar, se estima la entonación de ambas señales a través de un determina-

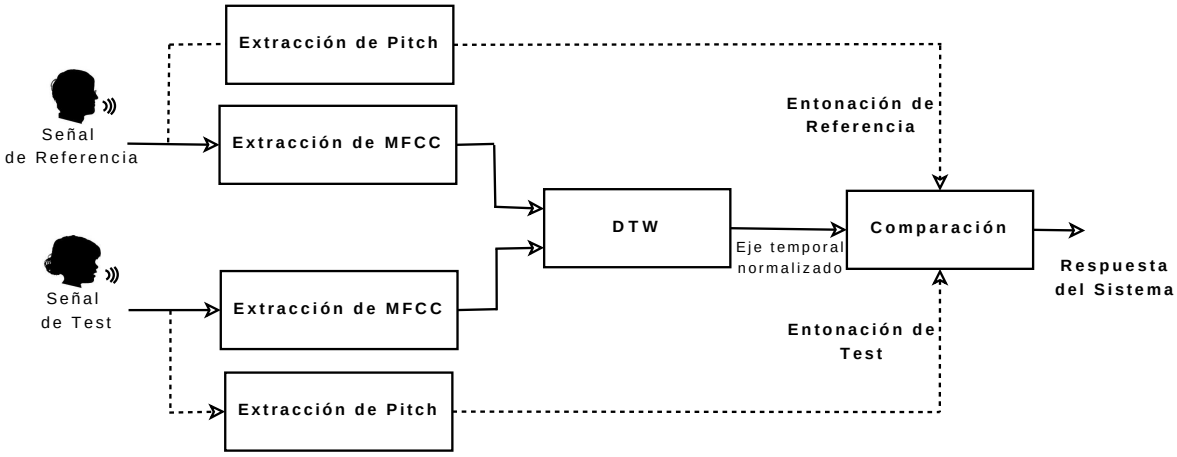

Figura 3.2: *Diagrama de bloques del sistema de evaluación de entonación.*

do algoritmo. Por otra parte, dichas señales son alineadas de acuerdo a sus coeficientes cepstrales en la escala de Mel (MFCC), usando el algoritmo de Alineamiento Temporal Dinámico DTW. Finalmente, una comparación de las curvas de *pitch* arroja un valor escalar, el cual se asocia a un puntaje o bien a una respuesta categórica de aceptación o rechazo, dependiendo si la entonación del estudiante se parece o no a la referencia. La ventaja de comparar la grabación de test con otra elocución y no con un modelo radica en que la información extralingüística de la entonación es fuertemente dependiente del contexto. Por ejemplo, en inglés normalmente las preguntas cuya resupesta es *si* o *no* poseen una entonación ascendente; no obstante, una persona molesta o aburrida no generará dicho patrón de entonación.

El alineamiento es de suma importancia para el sistema propuesto, ya que permite obtener una misma indexación temporal para las dos señales. Esto significa que el cuadro i de la referencia es comparable con el cuadro i de la señal de test. En términos simples, el uso de DTW compensa las diferencias temporales que puedan aparecer entre dos elocuciones generadas por dos personas distintas, como la velocidad del habla; la omisión e inserción de fonemas; silencios entre palabras de distinta duración; etc. Cabe

destacar que las señales a comparar son muy distintas entre si, ya que una es producida por un hablante nativo que además de proveer el patrón de entonación que se desea enseñar, pronuncia en forma correcta y habla con seguridad. Por otra parte, el usuario del sistema posiblemente se enfrente a éste con voz dubitativa, entrecortada y con una pronunciación errónea dada su condición de estudiante.

3.3.1. Pre-Procesamiento

Este preprocesamiento es aplicado tanto a las señales de referencia almacenadas como a las de test grabadas por el locutor. Primeramente, se eliminan componentes de frecuencias menores a $75[Hz]$ mediante un filtro pasa bajos, con el fin de reducir el ruido de fondo y del micrófono. Luego, se utiliza un detector de inicio y fin de señal útil, el cual tiene como objetivo eliminar espacios de silencio en los extremos de las elocuciones grabadas. Estos silencios no contienen información prosódica ni acústica de interés, y por lo tanto no deben ser considerados. Para cualquier sistema de procesamiento de voz, es de gran ayuda que las señales a analizar tengan una cantidad relativamente fija de silencio en los extremos.

La tarea de detectar el inicio y fin de señal no es sencilla. En presencia de ruido, el algoritmo podría confundirse e interpretar tal ruido como señal útil, o bien, si tal elocución fue grabada con un nivel muy bajo, es altamente probable perder cuadros con información relevante. En este trabajo, el detector de inicio y fin de señal útil toma sus decisiones a partir de umbrales de energía.

3.3.2. Alineamiento fonético

El algoritmo DTW es aplicado para alinear *frame* a *frame* la elocución del alumno con una referencia canónica de acuerdo a la similitud de su representación de MFCC. Para medir la distancia entre dos vectores de coeficientes, la alternativa más simple es la distancia euclidiana. Sin embargo, dado que cada uno de los MFCC y sus respectivas derivadas poseen varianzas distintas entre sí, es conveniente utilizar una medida de distancia que aminore este efecto. Para ello, se utiliza la distancia de *Mahalanobis*, definida por la ecuación 3.1:

$$d_{mahalanobis}(r,t) = \sqrt{(r-t)\Sigma^{-1}(r-t)^T} \tag{3.1}$$

En esta expresión, r es el vector de MFCC de referencia, y t el vector de test mientras que Σ es la matriz de covarianza, la cual es determinada a través de un conjunto de señales de entrenamiento, mediante un proceso iterativo que se muestra en la Figura 3.3.

Sean x_{1k} y x_{2k} las señales de voz de entrenamiento (con $k = 1 \dots K$) divididas en *frames* los cuales a su vez están representados por N_c coeficientes cepstrales. Para todo k, las respectivas señales x_{1k} y x_{2k} son alineadas con el algoritmo DTW usando la distancia de Mahalanobis. Dado que no se tiene la matriz de covarianza para la primera iteración, se toma $\Sigma = I$ (matriz identidad). De esta forma, la distancia expresada en la ecuación 3.1 se reduce a una distancia euclidiana clásica.

A partir del alineamiento se obtiene una matriz de distancias d_i (donde i representa la i-ésima iteración), cuyas dimensiones son N_c por la cantidad de *frames* que existen entre todos los pares de señales alineados. Luego, se calcula $\Sigma_{i+1} = d_i d_i^T$, matriz que resulta ser de $N_c \times N_c$. Si se cumple un cierto criterio de convergencia, se detiene el pro-

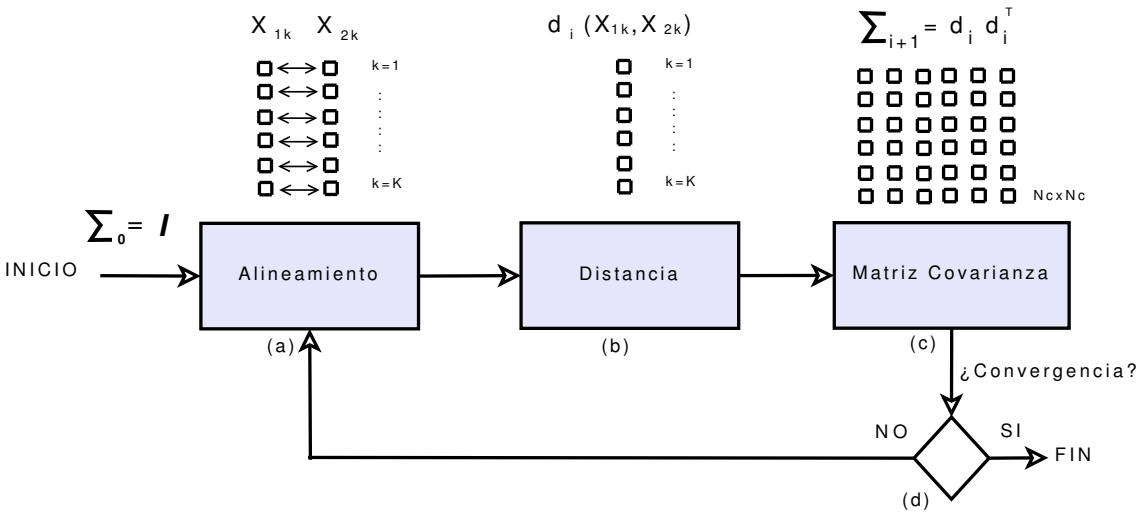

Figura 3.3: *Proceso iterativo para obtener matriz de covarianza* Σ.

ceso y Σ_{i+1} es la matriz buscada; en caso contrario se usa esta misma para la siguiente iteración. Como criterio de convergencia, se monitorean los cambios entre las matrices de covarianza resultantes entre dos iteraciones consecutivas: Si $|\Sigma_n - \Sigma_{n+1}| \leq \epsilon$, entonces el proceso se detiene.

3.3.3. Extracción de *pitch*

Se debe utilizar un método de detección de *pitch* confiable, ya que eventuales errores en esta etapa podrían alterar el desempeño del sistema. Una vez que la frecuencia fundamental f_0 es estimada para cada cuadro, se representa en semitonos de acuerdo con la ecuación 3.2.

$$p(k)_{semitonos} = 12 \log_2 p(k) \tag{3.2}$$

A esta curva $p(k)$ en semitonos se le aplica un proceso de interpolación lineal, el cual consiste en llenar los segmentos sordos, con el fin de que $p(k)$ quede definida para todos los *frames* de la señal. La Figura 3.4 muestra un ejemplo de interpolación lineal.

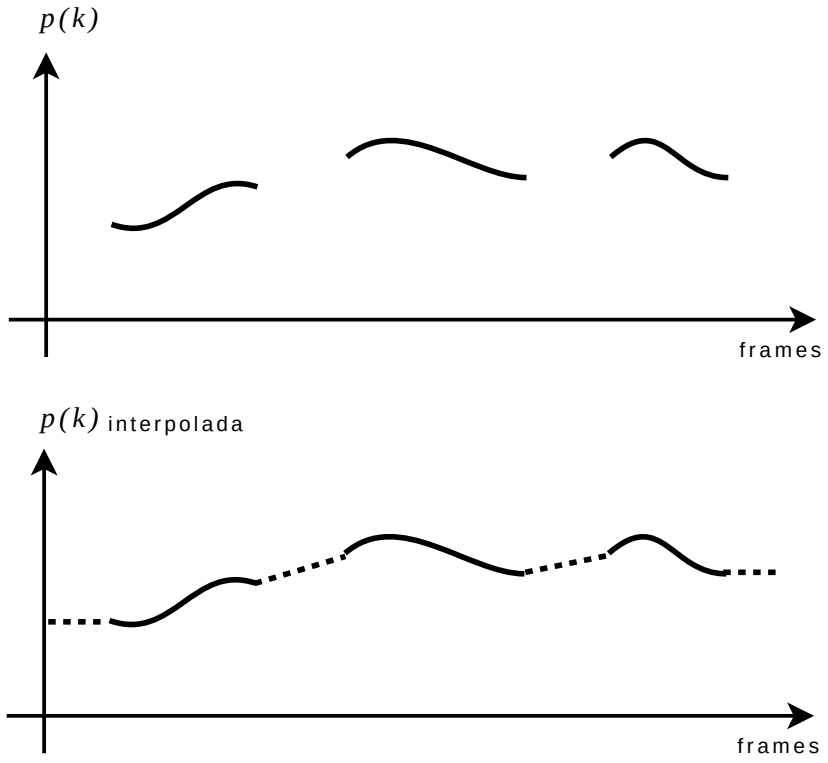

Figura 3.4: *Curva de pitch $p(k)$ (arriba), y $p(k)$ interpolada linealmente.*

La idea es unir todos los segmentos sonoros consecutivos mediante líneas rectas. Para ello, se traza una línea entre el último punto de un segmento sonoro con el primer punto del segmento siguiente. Para rellenar los *frames* iniciales, se toma el primer cuadro sonoro y se traza una recta con pendiente nula que pase por dicho punto; análogamente, para completar el tramo final de $p(k)$ se traza una recta con pendiente nula

que pase por el último *frame* sonoro.

Por último, se normaliza $p(k)$ sustrayendo su media aritmética $p(\bar{k})$. Observar que no es necesario obtener el valor exacto del pitch en cada punto, más bien se requiere calcular con precisión el contorno del *pitch*, esto es, la tendencia de f_0 en función del tiempo. Notar también que dada la metodología planteada, los errores de octava en la estimación de la frecuencia fundamental f_0 pueden alterar considerablemente la estimación de la entonación.

3.3.4. Comparación de curvas

3.3.4.1. Criterio de similitud

Una vez que se ha efectuado el alineamiento mediante DTW entre una señal de referencia y una de test; y además se ha calculado el *pitch*, el sistema está en condiciones de calcular la similitud de la entonación de ambas elocuciones. Dado un camino óptimo $i_R = \phi_R(k)$ y $i_T = \phi_T(k)$ que ha sido determinado mediante el algoritmo DTW, se define la *medida de similitud* $S(d, p)$ como:

$$S(d, p) = d(p_R(i_R), p_T(i_T)) \tag{3.3}$$

En esta ecuación, p_R y p_T son las curvas de pitch para la referencia y para la señal de test respectivamente, mientras que $d(\cdot, \cdot)$ es una función de distancia vectorial, la cual arroja un valor escalar que indica la similitud entre ambas curvas de *pitch*. Es posible elegir cualquier distancia, en el caso de este trabajo, se han escogido dos: la *correlación* y la *distancia euclideana*, definidas por:

$$d_{corr}(x,y) = \frac{\sum_{k=1}^{K}(x_k - \bar{x})(y_k - \bar{y})}{\sigma_x \sigma_y}$$
$$d_{euclidiana}(x,y) = \sqrt{\sum_{k=1}^{K}(x_k - y_k)^2} \tag{3.4}$$

La correlación es una medida estadística que tiene la ventaja de estar confinada en el intervalo $[-1, 1]$. A mayor d_{corr}, se tiene mayor similitud entre los vectores que se está comparando. Por otra parte, la distancia euclidiana puede arrojar cualquier valor mayor o igual a cero, luego es una medida no acotada. Al contrario de la correlación, la distancia euclidiana tiene un comportamiento inverso, ya que a mayor d_{euclid} mayor es la distancia entre dos vectores y por lo tanto su similitud es menor.

Dado el valor de la similitud, es posible entregar un puntaje o *score* en vez de una respuesta binaria, aplicando una transformación a la medida de similitud a una determinada escala de notas (por ejemplo de 1 a 5) o bien a una respuesta binaria como correcto/incorrecto. Incluso se podrían graficar los puntos de los contornos de pitch de referencia y test, para que el estudiante pueda visualizar y corregir sus errores. En efecto, se ha desarrollado una gran cantidad de programas educativos que dejan en evidencia las ventajas del uso de contornos de pitch como *feedback* visual (Botinis et al., 2001).

3.3.4.2. Uso de la Derivada de $p(k)$

Es posible obtener la primera derivada de las curvas de *pitch*, donde se extrae información de las subidas y bajas de la curva melódica. Ésta se calcula sobre $p(k)$

después de haber sido interpolada, mediante la recurrencia:

$$Dp(k) = p(k) - p(k-1) \tag{3.5}$$

De esta forma, en la ecuación 3.3 se puede reemplazar al *pitch* $p(k)$ por su derivada, resultando:

$$S(d, Dp) = d(Dp_R(i_R), Dp_T(i_T)) \tag{3.6}$$

En este caso Dp_R y Dp_T son las derivadas de las curvas de *pitch* de referencia y test respectivamente. Así, la derivada se puede utilizar como una característica alternativa a la entonación pura.

3.3.5. Evaluación de la Acentuación

La energía juega un rol muy importante en la acentuación. Como fue explicado en el capítulo 2, las sílabas enfatizadas o acentuadas tienden a ser pronunciadas con un *pitch* más elevado, pero además con mayor intensidad. El sistema de evaluación de acentuación que se propone en este trabajo se construye sobre el método de evaluación de entonación ya descrito. La idea es básicamente incluir la información de energía de las señales en la comparación, como muestra la Figura 3.5 (que es una extensión del diagrama mostrado en 3.2):

Se añade un bloque de extracción de energía, la cual se calcula para cada *frame* utilizando la ecuación 2.11 en decibeles, construyendo así la curva de energía $E(k)$. Observar que es necesario contar con una elocución de referencia generada por un hablante nativo, la cual debe estar correctamente acentuada. El bloque de comparación

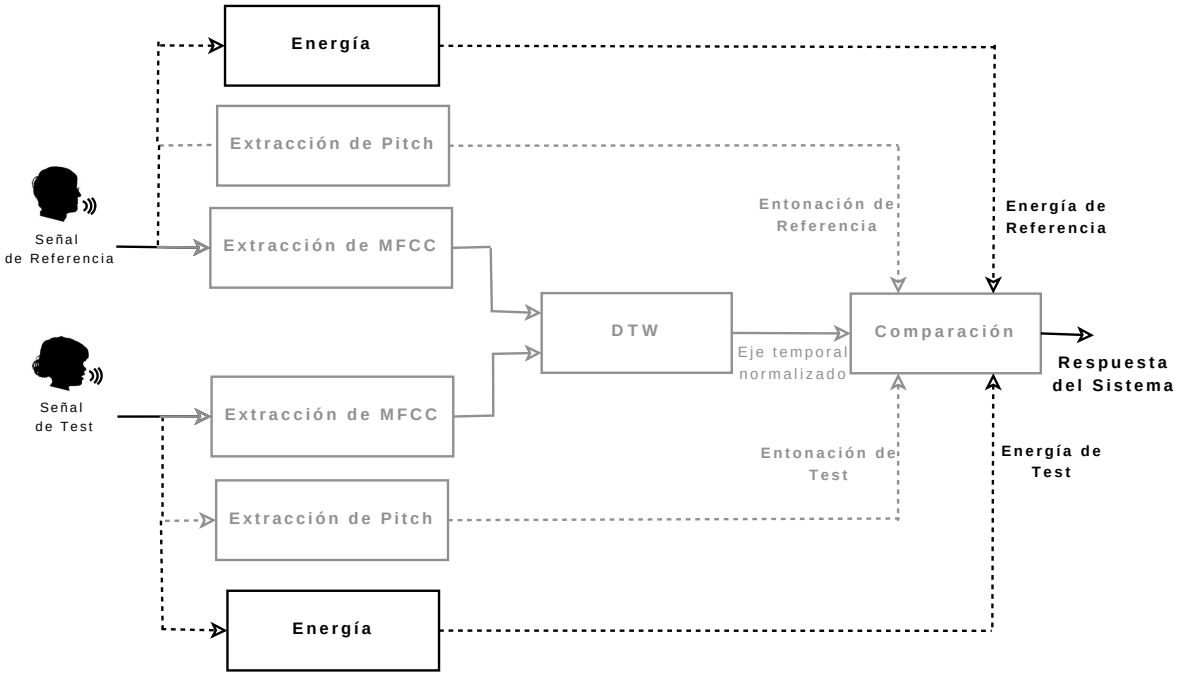

Figura 3.5: *Diagrama de bloques del sistema de evaluación de acentuación. Corresponde a una extensión del diagrama mostrado en la Figura 3.2.*

debe tomar una decisión en base a dos características, y no sólo una como ocurría en el caso de la entonación. Por lo tanto, se reformula la definición de similitud para el presente caso:

$$S(d,p,E) = \alpha \frac{S(d,p)}{\sigma_p} + (1-\alpha)\frac{S(d,E)}{\sigma_E} \tag{3.7}$$

La medida de similitud de acentuación se genera combinando linealmente las similitudes de energía y el contorno de *pitch*. Además, es necesario que ambos sumandos tengan la misma varianza, por lo que la similitudes son divididas por sus respectivas desviaciones estándar.

3.3.5.1. Decisión correcto/incorrecto

Una vez que se cuenta con un valor de similitud, es posible una decisión en base a éste. Si el *pitch* de referencia *se parece* (o *no se parece*) al de test, el sistema entrega una respuesta de aceptación mediante el valor 1 binario (o rechazo con un 0 binario). Para ello, se utiliza una función escalón:

$$u(S,\theta) = \begin{cases} 1 & \text{si } S \geq \theta \\ 0 & \text{en otro caso} \end{cases} \tag{3.8}$$

En esta ecuación, θ es un umbral de decisión para la similitud S, el cual establece el límite entre similitudes aceptadas o rechazadas. Este valor debe ser escogido cuidadosamente, ya que en definitiva es el parámetro que determina qué tan exacto es el sistema. La Figura 3.6 muestra dos sistemas de evaluación de entonación idénticos, pero con distintos umbrales de decisión θ. En el ejemplo, (a) es altamente permisivo, mientras que (b) es altamente exigente.

Un sistema permisivo tiende a aceptar como similares dos acentuaciones, incluso si no son muy parecidas. Este hecho es favorable ya que existe baja probabilidad de no aceptar que dos acentuaciones son similares dado que objetivamente lo son, pero a su vez es riesgoso aceptar como parecidas dos curvas de *pitch* que en realidad no lo son. Por el contrario, un sistema altamente exigente no acepta como similares acentuaciones que tienen pequeñas diferencias entre si, con lo cual la probabilidad de aceptar como similares dos acentuaciones objetivamente distintas es muy baja.

La elección de un umbral de decisión adecuado no es trivial, ya depende de la aplicación. Por ejemplo, la exigencia para un niño en edad escolar que recién comienza a

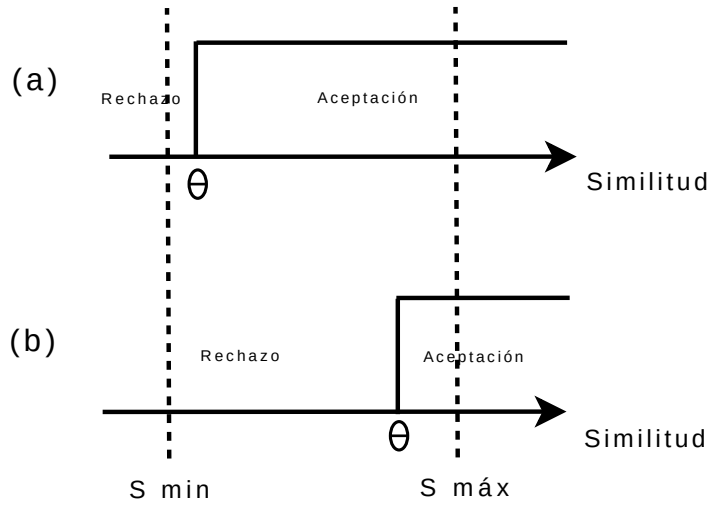

Figura 3.6: *Un sistema permisivo (a) y otro exigente (b).*

aprender una lengua extranjera es muy distinta a la de un ejecutivo que desea mejorar su inglés en presentaciones. Es claro que en el primer caso, la exigencia debe ser relativamente baja para no provocar frustración en el pequeño alumno, mientras que en el segundo caso es necesaria una alta exigencia para que el experimentado usuario adquiera características prosódicas muy parecidas a las que posee un hablante nativo.

3.4. Experimentos

3.4.1. Base de Datos

Está compuesta por los siguientes grupos:

- ENTONACIÓN: 6 frases pronunciadas con 4 patrones de entonación diferentes, grabados por 6 hablantes expertos, y 5 no expertos utilizando 3 micrófonos simultáneamente. Las frases son:

- *What's your name?*
- *My name is Peter.*
- *It's made of wood.*
- *It's terrible.*
- *It was too expensive.*
- *I tried both methods.*

- ACENTUACIÓN: 12 palabras: *machine*, *alone*, *under*, *husband*, *yesterday*, *innocence*, *important*, *excessive*, *melancholy*, *caterpillar*, *impossible* y *affirmative*. Todas fueron pronunciadas con distintas variaciones de acentuación, y grabados por 4 hablantes expertos y 2 no expertos, utilizando 3 micrófonos simultáneamente.

Cada una de las señales de la base de datos fue etiquetada con el respectivo patrón de entonación o acentuación con el cual fue pronunciado. De cada subconjunto, se elige un hablante experto como **referencia**, cuyas grabaciones son usadas como modelo para evaluar la entonación (o acentuación). Las grabaciones de los restantes hablantes expertos y de los no expertos son etiquetadas como **test**. De esta forma, la base de datos queda dividida en cuatro subconjuntos, como se muestra en la Tabla 3.1.

	Subconjunto	**# de señales**
Entonación	EN-R	24
	EN-T	720
Acentuación	AC-R	12
	AC-T	540

Tabla 3.1: *Subconjuntos de la base de datos.*

Los patrones de entonación usados son *Descendente Bajo* (DB); *Descendente Alto* (DA); *Ascendente Bajo* (AB); y *Ascendente Alto* (AA). Un ejemplo de cada uno se muestra en la Figura 3.7. Las variaciones de acentuación consisten en pronunciar cada palabra con el acento enfático en todas sus sílabas. Por ejemplo, la palabra "important" posee las variaciones: **im**portant, im**por**tant e impor**tant**, siendo correcta sólo la segunda.

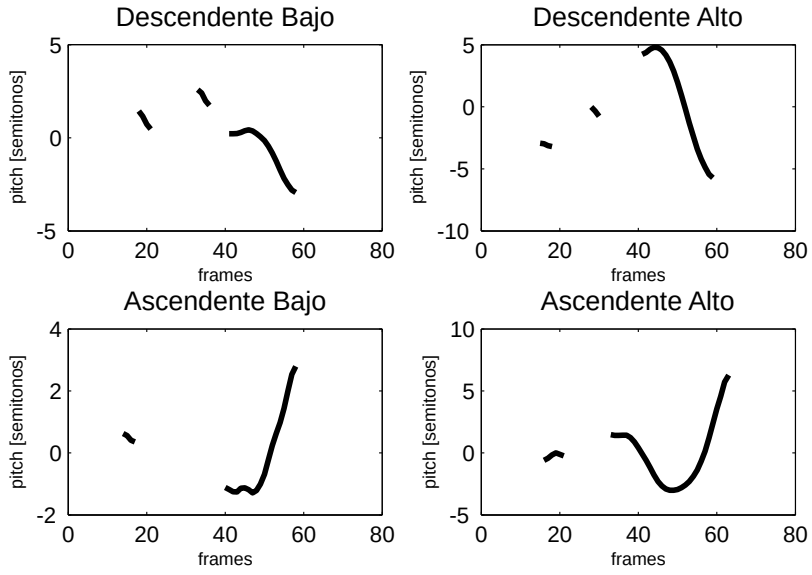

Figura 3.7: *Ejemplos de las cuatro variaciones de entonación utilizadas. La frase pronunciada es "What's your name".*

3.4.2. Condiciones Experimentales

Todas las señales fueron grabadas en el Laboratorio de Procesamiento y Transmisión de Voz (DIE, Universidad de Chile), a una frecuencia de muestreo f_s igual a $16000[Hz]$ y con un nivel de ruido similar al de una sala de clases. Los hablantes expertos poseen conocimientos de lingüística, por lo tanto, son capaces de reproducir cualquier patrón

de entonación o acentuación leyendo el texto etiquetado con una simbología adecuada. Por otra parte, los hablantes no expertos, quienes poseen un nivel intermedio de conocimiento del idioma inglés, deben oír una grabación de referencia y luego pronunciar la respectiva frase o palabra, tratando de generar una entonación (o acentuación) lo más parecida posible.

3.4.3. Configuración del sistema

Las señales son divididas en cuadros de 400 muestras, con un traslape de 200. Cada uno de los *frames* es representado por un vector con sus coeficientes cepstrales (MFCC), así como también los coeficientes dinámicos de primer y segundo orden, formando un total de 33 coeficientes.

Se utiliza el detector de pitch basado en autocorrelación exacta *Praat* (Boersma P. & Weenink D., 2008), el cual ha sido utilizado por varios trabajos recientes y por tanto, se considera en el estado del arte (Resh B. et al., 2007). La configuración de *Praat* es:

```
To Pitch (ac)... 0.0125 75 15 yes xx 0.45 0.01 0.35 0.14 600
```

El parámetro `xx` representa *silence threshold*, que corresponde al umbral de silencio que incide fuertemente en la decisión sonoro/sordo. El sistema está configurado para asignar dinámicamente este parámetro entre 0,01 y 0.3 en función de la energía de la señal, representada por el promedio de los 8 mínimos dentro de los 20 *frames* de mayor energía. Una vez que se ha calculado el *pitch*, se aplican condiciones de continuidad para eliminar saltos de octavas, y un filtrado mediano para eliminar puntos erróneos.

Adicionalmente, se implementó un detector de pitch llamado *lptvPitch*, el cual está basado en autocorrelación. La decisión sordo/sonoro se toma una vez que se ha calculado la autocorrelación para un *frame*, mediante un umbral de energía. Como etapa de pre-procesamiento, se aplica un filtro pasabanda con frecuencias de corte $75 - 600[Hz]$.

3.4.4. Experimentos de Alineamiento

En el presente trabajo, es muy importante que el Alineamiento Temporal Dinámico funcione en forma correcta, ya que imprecisiones en esta etapa podrían producir respuestas indeseadas o inesperadas. En virtud de lo anterior, es necesario implementar una maqueta de pruebas que permita aislar el comportamiento del algoritmo DTW, para así encontrar su configuración óptima y además mostrar que funciona adecuadamente. Los experimentos que se detallan a continuación intentan cumplir este objetivo.

3.4.4.1. Descripción del Experimento

Se desea medir la correctitud del método DTW. Dadas dos elocuciones, es posible estimar manualmente puntos de su alineamiento óptimo y compararlos con el camino obtenido con DTW a través de alguna medida de distancia. Con este valor se intenta cuantificar el funcionamiento del algoritmo para distintas configuraciones. Para efectuar esta prueba, se elige un grupo de 240 señales del subconjunto EN-T, compuestas por las grabaciones de 2 micrófonos de 3 hablantes expertos y 2 no expertos.

Las grabaciones son etiquetadas por personas, quienes identifican el inicio de cada sílaba dentro de la frase. Para esto, se asocian marcas a aquellos *frames* que sean considerados inicio de sílaba. En la Figura 3.8 se muestra un ejemplo, donde se puede ver

la elocución “What’s your name” que ha sido etiquetada en forma manual. Las líneas verticales corresponden a las marcas que identifican los comienzos de las palabras de una sílaba “What’s”; “your” y “name” dados por las muestras 2992; 5637 y 8016 respectivamente.

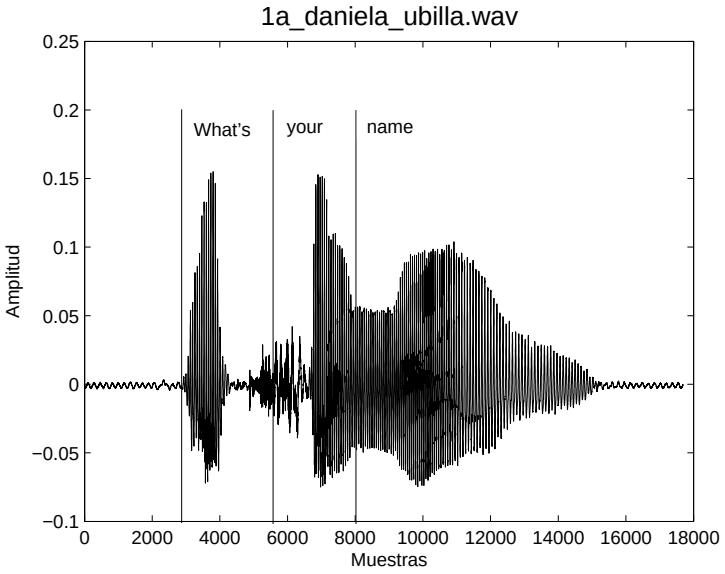

Figura 3.8: *Ejemplo de forma de onda etiquetada.*

Una vez que las señales han sido etiquetadas se ejecuta el algoritmo DTW, alineando todos los posibles pares de grabaciones donde se haya pronunciado la misma frase sin importar la prosodia, y que además tanto los locutores como los micrófonos sean distintos (en total 1116 pares de señales cumplen estas condiciones). El resultado de esto se puede representar en gráficos de dos dimensiones cuyos ejes corresponden a los *frames* de cada una de las señales alineadas, como el que se muestra en la Figura 3.9.

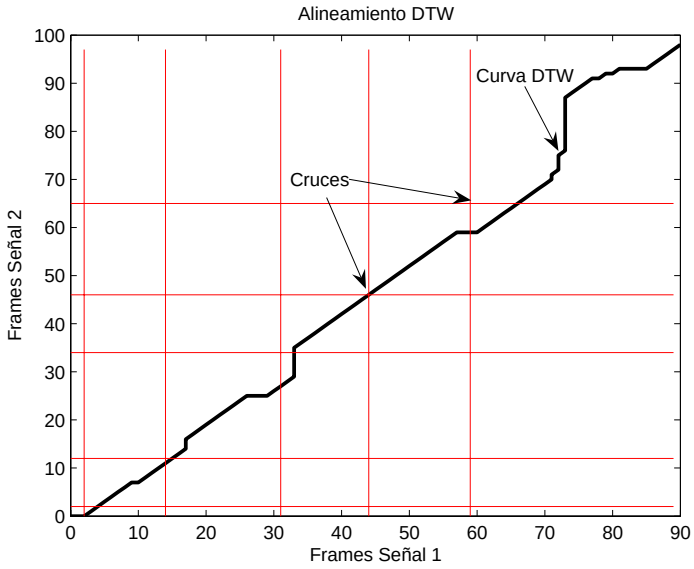

Figura 3.9: *Par de señales alineadas.*

Las líneas delgadas horizontales y verticales corresponden a las marcas de inicio de sílaba de cada elocución, las cuales se intersectan entre sí formando un cruce. Si el alineamiento fuera perfecto, entonces todos estos cruces debieran coincidir con el camino dado por DTW y por lo tanto, la correctitud del alineamiento puede calcularse midiendo la distancia que existe entre ellos.

El error de alineamiento para un cruce dado se puede calcular a través de una estimación de su distancia mínima a la curva DTW. Para ello, se utiliza una distancia dada por la expresión:

$$e_{alin} = \frac{1}{2}\sqrt{d_h^2 + d_v^2} \tag{3.9}$$

En la Figura 3.10 se muestra el significado de los valores d_h y d_v, que corresponden a las distancias horizontales y verticales entre el cruce y la curva. Además, d_{min} representa una *estimación* de la distancia mínima que se quiere calcular. El promedio de

todos los e_{alin} corresponde una cantidad escalar que representa una idea global de la correctitud del alineamiento DTW cuando se aplica a un determinado un par de señales.

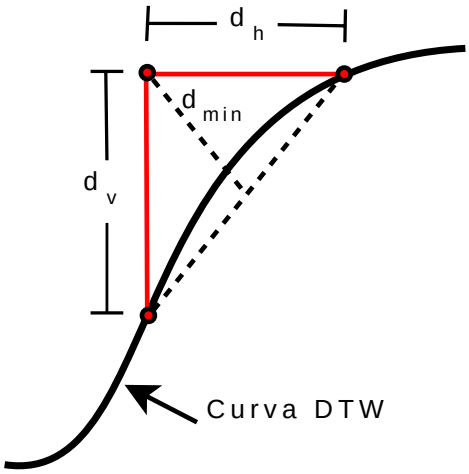

Figura 3.10: *Estimación de la distancia mínima entre un cruce y la curva DTW.*

Por otra parte, el error máximo que el alineamiento puede cometer es igual a la distancia entre los bordes de la banda de Sakoe & Chiba, es decir $r\sqrt{1+(\frac{T_y}{T_x})^2}$. Considerando que las señales a comparar tienen un número de cuadros similar entre sí ($T_x \approx T_y$), entonces el error máximo se resulta ser $r\sqrt{2}$. Así, e_{alin} puede estimarse relativo a este valor como porcentaje.

Para este trabajo, se ha implementado una versión de DTW que utiliza la banda de Sakoe & Chiba como restricción global, cuyo radio ha sido determinado por inspección (su valor es $r = 24$). La ecuación de programación dinámica 2.17 corresponde a la condición $P = 0$ (Sakoe & Chiba, 1978). El algoritmo se prueba utilizando dos configuraciones, las cuales difieren únicamente en la medida de distancia utilizada (euclidiana y Mahalanobis). Para determinar la matriz de covarianza Σ de la distancia de Mahalanobis se utiliza un subconjunto de 848 señales, tomadas de EN-T y AC-T.

3.4.4.2. Resultados

El sistema mostrado en este trabajo debe comparar señales que fueron capturadas con distintos micrófonos y generadas por distintos locutores. Además, es altamente probable que ambas posean distinta prosodia entre sí. En virtud de lo anterior, es necesario verificar que el comportamiento del algoritmo DTW no se ve afectado por estas diferencias. En la Tabla 3.2 se muestra el error de alineamiento DTW para dos configuraciones distintas. El conjunto de prueba ha sido dividido en dos categorías: igual prosodia y distinta prosodia.

	Prosodia	
	Igual	*Distinta*
Dist. euclidiana	4,78 %	4,94 %
Dist. Mahalanobis	4,30 %	4,45 %

Tabla 3.2: *Error de alineamiento del algoritmo DTW.*

3.4.5. Experimentos de Evaluación de Entonación

3.4.5.1. Descripción del Experimento

Se desea evaluar la calidad del sistema de evaluación de entonación para distintas configuraciones. Cada una de las señales del subconjunto EN-T es enfrentada con todas sus referencias posibles dentro de EN-R, generando un total de 3456 comparaciones, como las de los ejemplos mostrados en la Figura 3.11. En cada experimento, se utiliza la correlación y la distancia Euclidiana como funciones de comparación, y se aplican tanto sobre la entonación pura como sobre su derivada. De esta forma, se obtienen 4 medidas de similitud en cada uno de los casos: *Euclid(p)*; *Corr(p)*; *Euclid(Dp)* y *Corr(Dp)*.

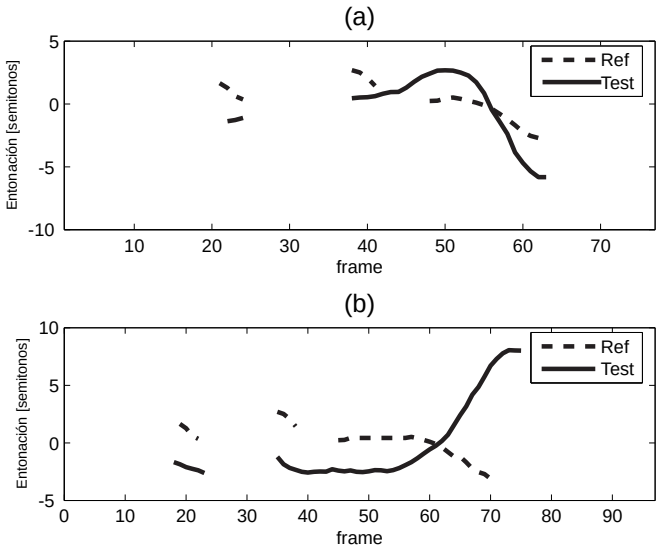

Figura 3.11: *Ejemplo de dos experimentos de evaluación de entonación: un caso de alta (a) y otro de baja similitud (b).*

Para evaluar el sistema, es necesario comparar la respuesta de éste con un puntaje objetivo que indique la similitud en cada experimento. Se puede estimar un puntaje a priori en cada experimento, ya que las señales de los subconjuntos EN-R y EN-T poseen etiquetas que indican el patrón de entonación con el cual fueron grabadas. A partir de éstos, se genera un valor que representa la similitud objetiva para cada comparación utilizando la Tabla 3.3. Si un usuario es capaz de reproducir exactamente la misma entonación, o bien si puede seguir la tendencia general del pitch a lo largo de la frase, debiera obtener una calificación alta. Por otra parte, se considera grave que el alumno produzca una entonación descendente cuando la referencia es ascendente o viceversa, y en estos casos se asigna una calificación baja.

Ciertamente, la asignación de notas mostrada en la Tabla 3.3 no es la única que se puede utilizar. La medida de similitud puede ser convertida en una respuesta categórica que indique si el estudiante ha producido una entonación similar a la referencia o no.

	DA	**DB**	**AA**	**AB**
DA	5	4	1	2
DB		5	2	1
AA			5	4
AB				5

Tabla 3.3: *Nota o score asociado a una comparación de entonación, usando un criterio no estricto.*

La Tabla 3.4 muestra el *score* que se asocia a cada comparación: 5 si se parecen y 1 en otro caso.

	DA	**DB**	**AA**	**AB**
DA	5	1	1	1
DB		5	1	1
AA			5	1
AB				5

Tabla 3.4: *Nota o score asociado a una comparación de entonación, usando un criterio estricto.*

3.4.5.2. Resultados

Para cada una de las medidas de similitud se calcula la correlación de acuerdo a las notas mostradas en las Tablas 3.3 y 3.4. Los resultados de esta prueba son presentados en la Tabla 3.5.

Los experimentos pueden ser diferenciados, de acuerdo al tipo de señales que están comparando. De esta forma, se muestran las correlaciones separando el subconjunto

	Escala	
	No estricta	*Estricta*
Corr(P)	0,87	0,55
Euclid(P)	0,62	0,40
Corr(DP)	0,79	0,49
Euclid(DP)	0,46	0,31

Tabla 3.5: *Correlación en evaluación de entonación para distintas medidas de similitud.*

EN-T en hablantes expertos y no expertos (Tabla 3.6). En la Figura 3.12 se muestra la correlación usando los tres micrófonos de dicho subconjunto.

	Hablante	
	Experto	*No-experto*
Corr(P)	0,88	0,85
Euclid(P)	0,62	0,61
Corr(DP)	0,77	0,79
Euclid(DP)	0,44	0,52

Tabla 3.6: *Correlación en evaluación de entonación, diferenciados en hablantes expertos y no-expertos. Se utiliza la escala de puntajes no estricta.*

Finalmente, se evalúa el comportamiento del sistema cuando se utilizan dos algoritmos de detección de *pitch* diferentes. Estos resultados son presentados en la Figura 3.13

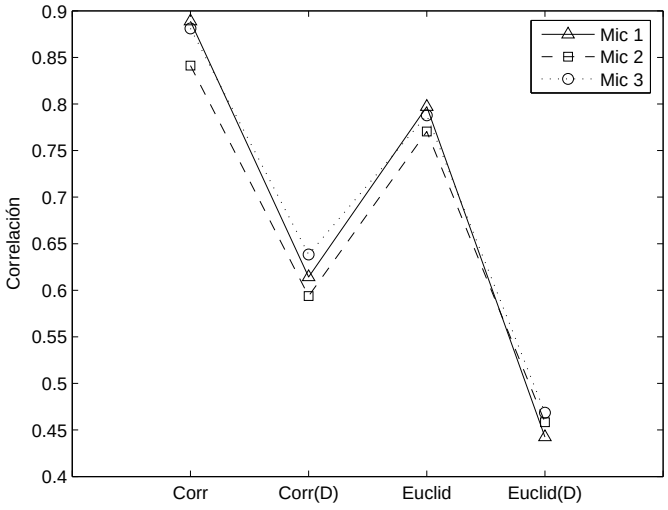

Figura 3.12: *Correlación en evaluación de entonación, diferenciados por micrófonos. M1 es de alta calidad, mientras que M2 y M3 son de bajo costo. Se utiliza la escala de puntajes no-estricta.*

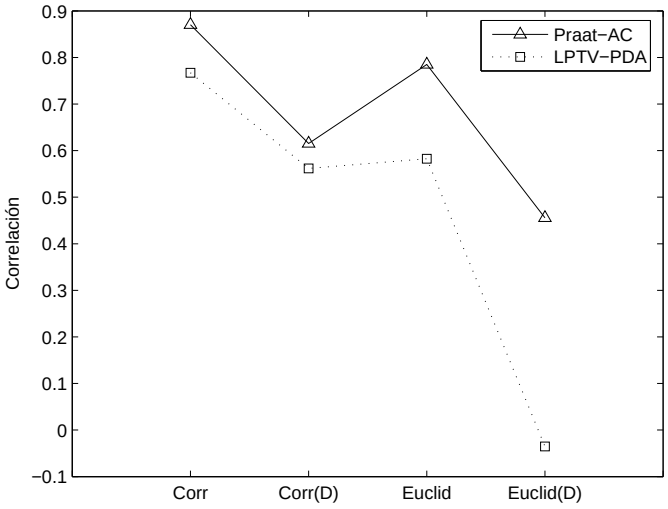

Figura 3.13: *Correlación en evaluación de entonación, para dos algoritmos de detección de pitch distintos: Praat-AC y LPTV-PDA. Se utiliza la escala de puntajes no-estricta.*

3.4.6. Experimentos de Evaluación de Acentuación

3.4.6.1. Descripción del Experimento

Cuando se evalúa la pronunciación o la entonación de un estudiante, es posible asignar puntajes a sus intentos de acuerdo a una escala de notas. Sin embargo, en un ejercicio de evaluación de acentuación la respuesta es categórica: correcto o incorrecto, y no tienen sentido puntajes intermedios. En consecuencia, el sistema puede determinar que:

- La acentuación es correcta dado que es correcta.
- La acentuación es incorrecta dado que es correcta.
- La acentuación es correcta dado que es incorrecta.
- La acentuación es incorrecta dado que es incorrecta..

Los dos primeros casos corresponden a respuestas correctas por parte del sistema de evaluación de entonación y son denominados *verdaderos-positivos* (VP) y *verdaderos-negativos* (VN) respectivamente. Los últimos dos casos son respuestas erróneas y se denominan *falsos-positivos* (FP) y *falsos-negativos* (FN). Estos valores pueden ser calculados como porcentajes o tasas, cuando se tiene una cantidad de comparaciones suficientemente grande, valores que varían de acuerdo al umbral de decisión θ utilizado.

Al graficar FN versus FP para distintos umbrales, se obtiene la curva de operación del receptor o simplemente curva ROC (*Receiver Operating Charasteristic*). El área bajo esta curva es un indicador de la habilidad discriminativa del sistema bajo el rango completo de umbrales de decisión en el que éste es probado: mientras menor sea esta área, mejor es el desempeño mostrado. El *Equal Error Rate* (EER) corresponde al punto de la curva donde FP se iguala a FN. Este valor es ampliamente utilizado para

medir el desempeño de sistemas biométricos.

El subconjunto AC-T es comparado con AC-R, lo cual genera un total de 540 experimentos. En esta prueba se utiliza el *pitch*, la energía y sus derivadas como medida de similitud.

3.4.6.2. Resultados

La Figura 3.14 muestra todas las posibles combinaciones. Finalmente, la Tabla 3.7 muestra un resumen de todas las configuraciones probadas para evaluación de acentuación.

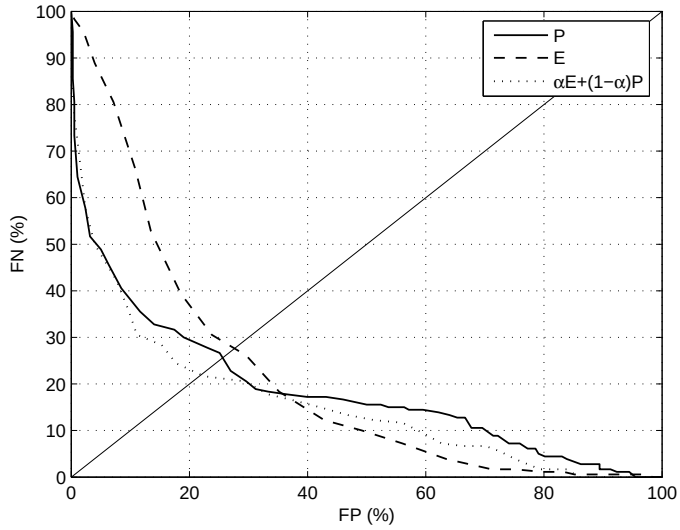

Figura 3.14: *Curvas ROC estimadas usando el pitch, la energía y la mejor combinación lineal de ambas (menor área ROC, $\alpha = 0,49$). Se ha utilizado la correlación como medida de similitud.*

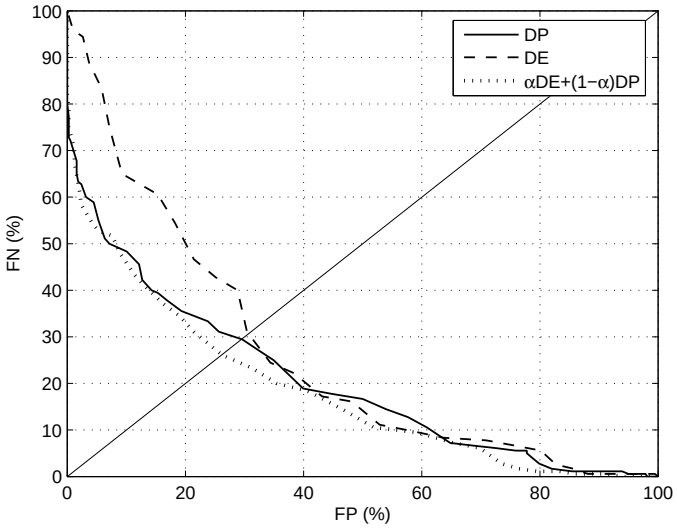

Figura 3.15: *Curvas ROC estimadas usando la primeras derivada pitch y la energía, y la mejor combinación lineal (menor área ROC, $\alpha = 0,66$). Se ha utilizado la correlación como medida de similitud.*

3.5. Discusión

De acuerdo a la Tabla 3.2, el error de alineamiento se mantiene prácticamente sin variación cuando el sistema tiene que comparar señales con igual o distinta prosodia. En efecto, el error aumenta en tan sólo en 0,16 % si se usa la distancia euclidiana, y un 0,15 % si se emplea Mahalanobis. Por lo tanto, este experimento muestra que la prosodia no altera el funcionamiento de DTW.

Por otra parte, se aprecia que el uso de la distancia de Mahalanobis en el alineamiento DTW genera una disminución relativa del error de alineamiento de 10 % y 9,9 % para experimentos de igual y distinta prosodia respectivamente. Esto se debe a que la normalización de la medida de distancia mediante la matriz de correlación compensa las diferencias de varianza de los coeficientes cepstrales, mejorando el desempeño del

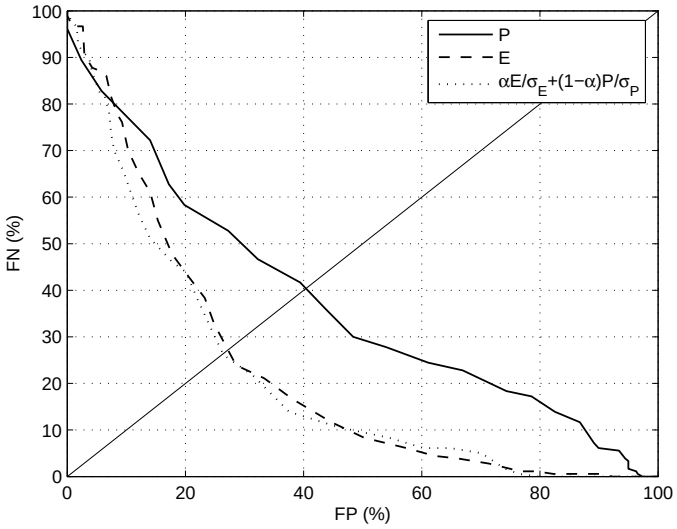

Figura 3.16: *Curvas ROC estimadas usando pitch, la energía y la mejor combinación lineal de ambas (menor área ROC, $\alpha = 0,23$). Se ha utilizado la distancia euclidiana como medida de similitud.*

algoritmo DTW.

La Tabla 3.5 muestra los resultados de los experimentos de evaluación de entonación. Se puede observar que al utilizar la correlación como medida de similitud se obtienen mejores resultados en comparación con la distancia Euclidiana, independiente de la escala de evaluación o medida de similitud usada. Además, se puede ver que el uso del *pitch* P siempre entrega mejores resultados que su derivada DP. En resumen, la mejor configuración del sistema $Corr(P)$, la cual entrega una correlación igual a 0,87 y 0,55 para las escalas no-estricta y estricta respectivamente.

En la Tabla 3.6 se muestra el desempeño del sistema cuando se evalúa la entonación de hablantes expertos y no-expertos. La diferencia entre estos tipos de hablantes es básicamente su calidad de pronunciación. El sistema debe ser robusto frente a pro-

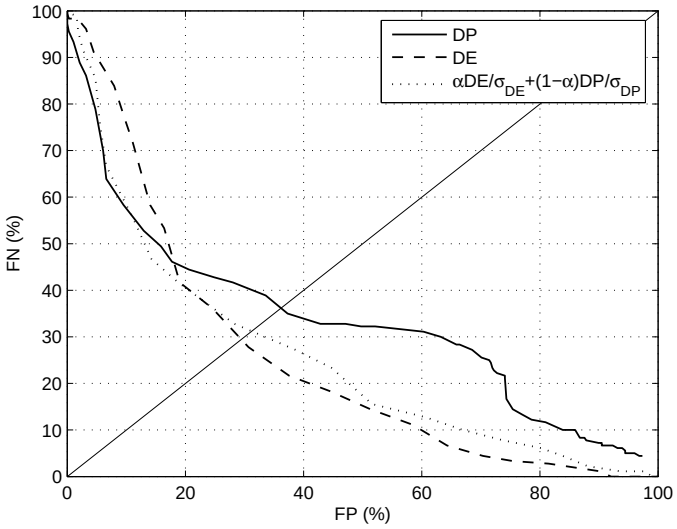

Figura 3.17: *Curvas ROC estimadas usando la primeras derivada pitch y la energía, y la mejor combinación lineal (menor área ROC, $\alpha = 0,77$). Se ha utilizado la distancia euclidiana como medida de similitud.*

nunciaciones imprecisas, ya que la evaluación de prosodia debe ser efectuada en forma aislada de la fonética. Se puede ver que una pronunciación inadecuada no incide mayormente en la evaluación de entonación, pues los coeficientes de correlación son similares para ambos tipos de hablantes. En efecto, para la mejor configuración ($Corr(P)$) la diferencia absoluta es de $0,03$.

De acuerdo con la Figura 3.12, el sistema no se ve mayormente afectado por la utilización de diferentes calidades de micrófonos. M1 es un aparato de alta calidad, mientras que M2 y M3 son de muy bajo costo, que introducen una cantidad importante de ruido a la señal de voz. Es claro que la variabilidad en los micrófonos es un problema tanto para la etapa de alineamiento como para la de extracción de f_0. Como se puede ver, las tres curvas están mínimamente desplazadas entre sí, lo cual muestra que el desempeño del método se mantiene invariante frente a los distintos micrófonos

	Área curva ROC	
	Correlación	*Distancia Euclid.*
E	0,184	0,371
P	0,212	0,222
$\alpha E + (1-\alpha)P$	0,156	0,210
DE	0,202	0,358
DP	0,257	0,249
$\alpha DE + (1-\alpha)DP$	0,181	0,253

Tabla 3.7: *Área bajo la curva ROC en evaluación de acentuación, para distintas medidas de similitud, usando distancia euclidiana y correlación.*

de la base de datos.

La Figura 3.13 muestra cómo el detector de *pitch* modifica el desempeño del sistema. Se puede observar que el coeficiente de correlación es mejor con *Praat* que con *LPTV-PDA*. Es evidente que el primer algoritmo es mucho mejor que el segundo, ya que tiene más tiempo de desarrollo. Este experimento más que mostrar que un algoritmo es mejor que otro, pretende mostrar que el sistema depende fuertemente del módulo de extracción de f_0, por lo que las mejoras sobre éste podrían incrementar el desempeño. Por ejemplo, se podría agregar una etapa de post-procesamiento para el *pitch*.

En las Figuras 3.14-3.17 se puede observar el desempeño del sistema de evaluación de acentuación usando la energía, el pitch y sus derivadas para distintas medidas de similitud. Se aprecia que la energía por sí sola discrimina mejor que el pitch para todos los casos. En un sistema de enseñanza de idioma, los usuarios tienden a generar una

	EER	
	Correlación	*Dist. Euclideana*
E	25,63 %	40,39 %
P	27,64 %	27,19 %
$\alpha E + (1-\alpha)P$	21,97 %	26,36 %
DE	29,50 %	36,18 %
DP	30,68 %	31,21 %
$\alpha DE + (1-\alpha)DP$	26,15 %	29,46 %

Tabla 3.8: *Equal Error Rate (EER) en evaluación de acentuación, para distintas medidas de similitud, usando distancia euclidiana y correlación.*

entonación plana e incluso ascendente, debido a su actitud vacilante al pronunciar palabras en un idioma distinto al nativo. En este experimento, los hablantes no-expertos representan a este tipo de usuarios, dadas las condiciones de grabación de la base de datos. En conclusión, el *pitch* es modificado por las emociones del hablante, lo cual explica que la energía discrimina mejor las diferencias de acento.

Como se esperaba, una combinación de ambas características siempre entrega mejores resultados. Como se puede observar en la Tabla 3.7, el área bajo la curva ROC es siempre menor para los experimentos que involucran el uso de energía y *pitch* simultáneamente. Además, el EER también es menor en estos casos como lo muestra la Tabla 3.8.

La mejor configuración para evaluación de *stress* se obtiene usando una combinación lineal de P y E, alcanzando un 21,97 % de EER, lo cual se traduce en que el sistema evalúa correctamente la entonación el 78,03 % de las veces. Al igual que en

evaluación de entonación, la correlación se comporta mejor que la distancia Euclidiana como medida de similitud. Finalmente, de las curvas y tablas se puede inferir que utilizar la derivada de las características no conduce a mejoras en los resultados.

Capítulo 4

Conclusiones

4.1. Análisis final

En este trabajo, se ha propuesto un método de evaluación de entonación, el cual estima las diferencias prosódicas entre dos señales. En el contexto de la enseñanza de segundo idioma, los elementos suprasegmentales de una palabra u oración pueden variar a causa de diversos factores, entre ellos la actitud, la intención y las emociones del hablante. Por esta razón, el sistema compara la señal de voz del usuario con una referencia pregrabada, y no con un modelo.

El sistema se implementó utilizando una versión mejorada del algoritmo DTW que utiliza la distancia de Mahalanobis para compensar las diferencias de varianzas de los coeficientes cepstrales. Los experimentos de alineamiento demostraron que el uso de esta distancia disminuye el error en un 10 % en comparación con el caso base (distancia euclidiana). Además, se mostró que el alineamiento es independiente de las diferencias prosódicas que eventualmente pueden existir entre las señales de referencia y de prueba.

El coeficiente de correlación entre los puntajes obtenidos en los experimentos y los

resultados esperados es igual a 0,87 para la mejor configuración. Ésta utiliza el *pitch* como característica, y la correlación como medida de similitud. A pesar de no tener un sistema base para comparar los resultados de evaluación de entonación, se puede concluir que la capacidad de comparación prosódica del método es bastante elevada. Por otra parte, el sistema demostró funcionar correctamente con micrófonos de baja calidad y también se probó su robustez frente a pronunciaciones incorrectas por parte de los usuarios (hablantes no expertos del subconjunto EN-T).

El método fue modificado para extender su funcionalidad y ser utilizado en evaluación de acentuación. Se obtiene un EER igual a 21,97 % para la configuración óptima, que utiliza la correlación como medida de similitud. Los experimentos han demostrado además que la energía en conjunto con el *pitch* entregan un mejor desempeño que si se utilizan dichas características en forma separada.

El aporte de este trabajo no radica en el alineamiento DTW ni métodos de extracción de características prosódicas. Como se puede observar, cada uno de los módulos que componen el sistema presentado en esta memoria son bastante conocidos y existen muchos trabajos sobre ellos. La contribución de este trabajo es el uso conjunto de tales módulos aplicado a la enseñanza de segundo idioma.

Si bien la técnica presentada fue probada para el inglés usando señales de hablantes nativos del español, el uso del sistema no está restringido a estos idiomas en particular. En efecto, podría ser utilizado en cualquier dialecto en la cual la entonación o la acentuación jueguen un rol significativo en el lenguaje y la comunicación.

El sistema propuesto puede ser combinado con un módulo de evaluación de pronunciación. Existen otros trabajos que incorporan la prosodia como parte de la calidad de

pronunciación (Dong B. et al, 2004). Cuantificar la similitud prosódica de dos señales puede tener muchas aplicaciones. Se puede implementar por ejemplo un sistema de enseñanza de tonos para idiomas como el chino mandarín. El método propuesto en este trabajo puede incluso ser utilizado en aplicaciones no relacionadas con la enseñanza de una lengua extranjera, como por ejemplo la estimación de discapacidades en el habla.

4.2. Trabajo futuro

Algunos módulos de la técnica de evaluación de entonación y *stress* presentada pueden ser modificados con el fin de obtener mejores resultados. Tal como se demostró en los experimentos, la correcta estimación de la frecuencia fundamental es de suma importancia en este sistema. El problema de detección de *pitch* es un tema que aún es materia de investigación, y si bien existen métodos que han alcanzado un nivel muy alto de exactitud, problemas como el ruido o el canal reducen considerablemente el desempeño de los algoritmos de detección de f_0. Como trabajo futuro se propone tratar estos problemas, para lo cual existen varias soluciones: incluir un módulo de post-procesamiento más robusto, aplicar alguna técnica de cancelación de ruido o de canal, o simplemente buscar una mejor forma de estimar f_0.

Si bien los resultados para evaluación de entonación son satisfactorios, hace falta comparar la respuesta del sistema con evaluaciones subjetivas de expertos. Por útlimo, es evidente que la exactitud en evaluación de acento mejoraría bastante si de alguna forma el sistema manejara información de la duración de cada segmento.

Referencias

Alkulaibi A., Soraghan J.J., Durrani T.S. (1996). Fast HOS Based Simultaneous Voiced/Unvoiced Detection and Pitch Estimation Using 3-Level Binary Speech Signals. 8th IEEE Signal Processing Workshop on Statistical Signal and Array Processing SSAP'96, pp. 194-197.

Austin, J. (1962). How to do things with words. Cambridge, Mass. Harvard University Press.

Bagshaw P.C. (1994). Automatic prosodic analysis for computer aided pronunciation teaching. PhD Thesis, The University of Edinburgh.

Boersma P., Weenink D. (2008) Praat: doing phonetics by computer (Version 5.0.29) [Computer program]. Retrieved July 14, 2008, from http://www.praat.org/

Botinis A., Granström B., Möbius B. (2001). Developments and paradigms in intonation research. Speech Communication, Volume 33, Issue 4, March 2001, pp. 263-296.

Brenier J.M., Cer D.M., Jurafsky D. (2005). The detection of emphatic words using acoustic and lexical features. INTERSPEECH-2005, pp. 3297-3300.

Brosnahan L.F., Malmberg B. (1970). Introduction to Phonetics, Cambridge, W.Heffer & Sons.

Brown J., Miller P. (2006). Automatic classification of killer whale vocalizations using dynamic time warping. Journal of the Acoustical Society of America, Aug. 2007, pp. 1201-1207.

Brown M., Rabiner L. (1982). An adaptive, ordered, graph search technique for dynamic time warping for isolated word recognition. Acoustics, Speech, and Signal Processing, Aug 1982, Vol. 30, Issue 4, pp. 535-544.

Chapelle C. (1997). Call in the year 2000: Still in search of research paradigms?. Language Learning & Technology, Vol. 1, No. 1, July 1997, pp. 19-43.

Chomsky, N. y Halle M. (1968). The Sound Pattern of English. New York, Harper & Row.

Criel J., Tsiporkova E. (2006). Gene Time Echipression Warper: a tool for alignment, template matching and visualization of gene expression time series. Bioinformatics 22, Vol. 2, pp. 251-252.

de Cheveigné, A., and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. Journal Acoustic Society Am., IRCAM.

Delmonte R., Peterea M., Bacalu C. (1997). SLIM: Prosodic Module for Learning Activities in a Foreign Language. Proc. ESCA, Eurospeech 97, Rhodes, Vol. 2,

pp. 669-672.

Dong B., Zhao Q., Zhang J., Yan Y. (2004). Automatic assessment of pronunciation quality. International Symposium on Chinese Spoken Language Processing, Dec. 2004, pp. 137-140.

Ehsani F., Knodt E. (1998). Speech technology in computer-aided language learning: Strengths and limitations od a new CALL paradigm. Language Learning & Technology, Vol. 2, No. 1, July 1998, pp. 45-60.

Face, T. (2006). Narrow Focus Intonation in Castillian Spanish Absolute Negatives. Journal of Language and Linguistics, Vol. 5, pp. 295-311.

Faundez-Zanuy M. (2007). On-line signature recognition based on VQ-DTW. Pattern Recognition archive Vol. 40, Issue 3, March 2007, pp. 981-992.

Franco H., Neumeyer L., Kim Y., Ronen O., (1997). Automatic pronunciation scoring for language instruction. ICASSP'97, 1997, Vol. 2, pp. 1471-1474.

Gavrila D.M., Davis L.S. (1995). Towards 3-D model-based tracking and recognition of human movement: a multi-view approach. Int. Workshop on Face and Gesture Recognition, pp. 272-277.

Gollmer K., Posten C. (1995). Detection of distorted pattern using dynamic time warping algorithm and application for supervision of bioprocesses. OnLine Fault Detection and Supervision in Chemical Process Industries.

Gu L., Harris J. (2003). SLAP: a system for the detection and correction of pronunciation for second language acquisition. International Symposium on Circuits and Systems, ISCAS '03, Vol. 2, pp. 580-583.

Hallyday, M. (1994). An Introduction to functional grammar (2nd edn.).

Hirose K., Fujasaki H., Seto S. (1992). A scheme for pitch extraction of speech using autocorrelationfunction with frame length proportional to the time lag. Acoustics, Speech, and Signal Processing, ICASSP-92, pp. 149-152.

Howard I. (1991). Speech Fundamental Period Estimation Using Pattern Classification, PhD Thesis. Faculty of Science, Phonetics & Linguistics, UCL, University of London.

Hung W. (2002). Use of fuzzy weighted autocorrelation function for pitch extraction from noisy speech. Electronic Letters 38, Volume 38, Issue 19, pp. 1148-1150.

I. S. Burnett and P. M. B. Gambino (1996). Pitch detection based on prototype waveforms. Proceedings of the ISSPA 96. Proceedings of Fourth International Symposium on Signal Processing and its Applications, Gold Coast, Qld., Australia.

Jones D., (1962). An Outline of English Phonetics, Cambridge: W.Heffer & Sons Ltd.

Keogh E., Pazzani M. (1999). Scaling up Dynamic Time Warping to Massive Datasets. 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'99), pp. 1-11.

Kim Ch., Sung W. (2004). Implementation of an intonational quality assessment system for a handheld device. INTERSPEECH-2004, pp. 1857-1860.

Kobakate H. (1987). Optimization of Voiced/Unvoiced Decisions in Nonstationary Noise Environments. , Hidefumi Kobatake, vol. No. 1, pp. 9-18, Jan. 1987, IEEE.

Legrand B., Chang C.S., Ong S.H., Neo S., Palanisamy N. (2008). Chromosome term classification using dynamic time warping. Pattern Recognition Letters, Volume 29, Issue 3, 1 February 2008, pp. 215-222.

MacAulay R. (1978). Maximum likelihood pitch estimation using state-variable techniques. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '78. Publication Date: Apr 1978, vol. 3, pp. 12-14.

Marchand S. (2001). An efficient pitch-tracking algorithm using a combination of Fourier transforms. Proceedings of DAFX-01, Limerick, Ireland.

Neumeyer L., Franco H., Digalakis V., Weintraub M. (2000). Automatic scoring of pronunciation quality. Speech Communication, Vol. 30, Issues 2-3, pp. 83-94.

Oh K., Un C. (1984). A performance comparison of pitch extraction algorithms for noisy speech. Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing, pp. 1-4.

Picone J. (1993). Signal Modeling Techniques in Speech Recognition. Proceedings of the IEEE, Vol. 81, No. 9, pp. 1215-1247.

Rabiner L.R. (1977). On the use of autocorrelation analysis for pitch detection.

IEEE Transactions Acoustic Speech, Signal Processing, vol. ASSP-25, no. 1, pp 24-33.

Rabiner L.R., Sambur M.R., Schmidt C.E. (1975). Application Of a nonlinear smoothing algorithm to speech processing. IEEE Trans. Speech and Audio Process. Vol. ASSP-23, pp. 552-555.

Ramírez D., Romero J. (2005). The pragmatic function of intonation in L2 discourse: English tag questions used by Spanish speakers. Intercultural Pragmatics. Vol. 2, 06/2005, pp. 151-168.

Resch B., Nilsson M., Ekman A., Kleijn W.B. (2007). Estimation of the Instantaneous Pitch of Speech. IEEE Transactions on Audio, Speech, and Language Processing. Volume 15, Issue 3, pp. 813-822.

Sakoe H., Chiba S. (1978). Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Signal Process. ASSP-26, pp. 43-49.

Shankera A.P., Rajagopalan A.N.(2007) Off-line signature verification using DTW. Pattern Recognition Letters, Vol. 28, Issue 12, 1 September 2007, pp. 1407-1414.

Shimamura T., Kobayashi H. (2001). Weighted autocorrelation for pitch extraction of noisy speech. IEEE Trans. Speech and Audio Process. Vol. 9, pp. 727-730.

Silverman H., Morgan D. (1990). The application of dynamic programming to connected speech recognition. IEEE ASSP Magazine, Vol. 7, no. 3, pp. 6-25.

Tepperman J., Narayanan S. (2005) Automatic Syllable Stress Detection Using

Prosodic Features for Pronunciation Evaluation of Language Learners. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. (ICASSP '05), Vol. 1, pp. 937-940.

Traynor P.L. (2003). Effects of Computer-Assisted-Instruction on different learners. Instructional psychology journal, pp. 137-143.

Tuzcu V., Nas S.(2005). Dynamic time warping as a novel tool in pattern recognition of ECG changes in heart rhythm disturbances. International Conference on Systems, Man and Cybernetics, 2005 IEEE, 12 Oct. 2005, Vol. 1, pp. 182-186.

Vivanco H. (1998-1999). Análisis fonético acústico de una pronunciación de "ch" en jóvenes del estrato medio-alto y alto de Santiago de Chile. Boletín de Filología de la Universidad de Chile, pp. 1257-1269.

Warschauer, M. (1996). CALL for the 21st Century. IATEFL and ESADE Conference, Barcelona, España.

Yong Duk Cho K., Al-Naimi K., Kondoz A. (2002). Pitch post-processing technique based on robust statistics. IEEE Electronics Letters, Vol. 38, 26 Sep 2002, pp. 1233-1234.

Zhao X., O'Shaughnessy D., Minh-Quang N. (2007). A Processing Method for Pitch Smoothing Based on Autocorrelation and Cepstral F0 Detection Approaches. Signals, Systems and Electronics, 2007. ISSSE '07, pp. 59-62.