



UNIVERSIDAD DE CHILE

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS

DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

EVALUACIÓN AUTOMÁTICA DE PROSODIA CON APLICACIONES EN  
ENSEÑANZA DE IDIOMAS Y DETECCIÓN DE EMOCIONES

TESIS PARA OPTAR AL GRADO DE DOCTOR EN INGENIERÍA ELÉCTRICA

JUAN PABLO ARIAS APARICIO

PROFESOR GUÍA:

NÉSTOR BECERRA YOMA

MIEMBROS DE LA COMISIÓN:

PATRICIO MENA MENA

ISMAEL SOTO GÓMEZ

CARLOS BUSSO RECARREN

Este trabajo ha sido parcialmente financiado por el Programa de Becas para estudios de Doctorado  
año 2008 de CONICYT y por los proyectos Fondef D05I10243 y Fondecyt 1070382/1100195.

SANTIAGO DE CHILE

JUNIO DE 2012

# Evaluación automática de prosodia con aplicaciones en enseñanza de idiomas y detección de emociones

Resumen de la Tesis para optar al grado de Doctor en Ingeniería Eléctrica

Juan Pablo Arias Aparicio

Profesor guía: Néstor Becerra Yoma

Santiago de Chile, Enero de 2013.

La prosodia es uno de los aspectos más importantes en la comunicación humana. La entonación, el ritmo, la intensidad y la duración entregan al locutor características como naturalidad, fluidez, intención, actitud, significado e incluso emoción. Por tanto, modelar y analizar la prosodia no sólo es interesante para el estudio del habla desde una perspectiva teórica, sino que también para las tecnologías de voz. En virtud de la creciente necesidad de interfaces hombre-máquina más parecidas a las interacciones humanas reales, los sistemas de procesamiento de patrones acústicos deben ser capaces de analizar e interpretar las características prosódicas.

En esta tesis se abordan dos problemas que involucran la modelación prosódica en señales de voz. En primer lugar, se presenta una técnica para la evaluación de la entonación en enseñanza de segundo idioma basado en un esquema *top-down*. El método propuesto separa la evaluación de entonación de la pronunciación a nivel de sonidos individuales. Dada una señal de referencia, el usuario puede escuchar y repetir una elocución dada imitando el patrón de entonación de referencia. La técnica estima una medida de similitud entre la señal de referencia y de test. Basado en este mismo esquema, se presenta un sistema para medir el acento léxico a nivel de sílabas usando la información de la frecuencia fundamental en conjunto con la energía. La técnica propuesta es independiente del texto y del idioma y minimiza el efecto de la calidad de pronunciación a nivel de segmentos.

Como resultado del esquema propuesto para enseñanza de idiomas, se presenta una estrategia para detectar emociones en señales acústicas usando modelos de referencia emocionalmente neutros. Primero, se considera un caso ideal léxico dependiente donde la referencia corresponde a una única señal. Luego, se construyen modelos de referencia léxico independientes usando una familia de contornos de F0. Para ello, se presenta un esquema novedoso basado en *functional data analysis* donde los modelos neutros se representan mediante una base de funciones y el F0 de test se caracteriza por las proyecciones sobre esta base. Finalmente, la técnica se extiende a nivel de sub-oración para detectar los segmentos que son emocionalmente más relevantes.

El método propuesto para evaluación de entonación entrega una correlación de evaluaciones subjetivos (dada por expertos) y objetivos (entregados por el sistema) igual a 0,88. El método para acento léxico entrega un *equal error rate* (EER) igual a 21,5%, que a su vez es comparable con las tasas de error entregadas por las técnicas de evaluación de pronunciación a nivel de segmento. Estos resultados sugieren que ambos sistemas pueden ser eficazmente usados en aplicaciones reales. Por su parte, el método de detección de emociones permite obtener una exactitud igual a 75,8% en la tarea de clasificación de neutro versus emocional en una base de datos actuada, que a su vez es 6,2% superior a la exactitud alcanzada por un sistema en el estado del arte. El sistema además se valida con una base de datos real, cuyos resultados muestran que el método propuesto puede ser utilizado en aplicaciones reales de detección de emociones.

*... Dedicado a Natalia*

# Agradecimientos

Quisiera agradecer a mi familia por su apoyo entregado durante estos años. Muchas gracias por su confianza, cariño, dedicación y comprensión. Agradezco a mis amigos institutanos por su incondicional y valiosa ayuda.

También quiero expresar mi agradecimiento al profesor Néstor Becerra Yoma por el conocimiento y la experiencia transmitidos y por haberme dado la posibilidad de realizar mis estudios de doctorado en el LPTV. Infinitas gracias a los compañeros y ex-compañeros del LPTV por toda la ayuda brindada.

Quiero agradecer especialmente al profesor Carlos Busso por haberme dado la oportunidad de trabajar como Visiting Scholar en el laboratorio MSP Multimodal Signal Processing (MSP) Lab, The University of Texas at Dallas, Texas, Estados Unidos, y por haber compartido conmigo su tiempo y experiencia.

Esta tesis fue financiada por el Programa de Becas para estudios de Doctorado año 2008 de la Comisión Nacional de Investigación Científica y Tecnológica (CONICYT) del Gobierno de Chile. El trabajo también contó con el apoyo de la beca del proyecto MECESUP FSM0601 para realizar una pasantía en el laboratorio MSP, The University of Texas at Dallas, TX, Estados Unidos. Además, parte de este trabajo fue financiado por los proyectos Fondef D05I10243 y Fondecyt 1070382/1100195.

# Tabla de Contenido

<b>1. Introducción</b>	<b>16</b>
1.1. Modelación prosódica . . . . .	16
1.2. Entonación y acento en enseñanza de idiomas . . . . .	18
1.3. F0 y emociones . . . . .	19
1.4. Objetivos . . . . .	20
1.4.1. Objetivo general . . . . .	20
1.4.2. Objetivos específicos . . . . .	20
1.5. Estructura de la tesis . . . . .	21
1.6. Contribuciones de la tesis . . . . .	23
<b>2. Revisión bibliográfica</b>	<b>25</b>
2.1. Introducción . . . . .	25
2.2. La voz humana . . . . .	27
2.3. Características segmentales . . . . .	29
2.4. Características suprasegmentales . . . . .	31

2.4.1.	Entonación . . . . .	31
2.4.2.	Acento . . . . .	34
2.4.3.	Duración . . . . .	35
2.4.4.	Otras características prosódicas . . . . .	35
2.5.	Algoritmos y técnicas utilizadas en procesamiento de voz . . . . .	36
2.5.1.	Parametrización acústica . . . . .	36
2.5.2.	Extracción de F0 . . . . .	37
2.5.3.	El algoritmo DTW . . . . .	43
2.5.4.	<i>Functional Data Analysis</i> (FDA) . . . . .	48
2.6.	Enseñanza de idiomas asistida por computador . . . . .	53
2.6.1.	Evaluación automática de prosodia . . . . .	56
2.6.2.	Medidas de desempeño en CALL . . . . .	58
2.7.	Reconocimiento y detección de emociones en señales de voz . . . . .	60
2.7.1.	Antecedentes . . . . .	60
2.7.2.	Representación de las emociones . . . . .	62
2.7.3.	Técnicas estándar en reconocimiento de emociones en señales de voz . . . . .	63
2.7.4.	Avances recientes . . . . .	65
2.7.5.	Medidas de desempeño en reconocimiento y detección de emociones	66
<b>3.</b>	<b>Evaluación de entonación y acento en enseñanza de idiomas</b>	<b>69</b>
3.1.	Introducción . . . . .	70

3.2.	La importancia de la entonación en enseñanza de segundo idioma . . . .	75
3.2.1.	Definiciones . . . . .	75
3.2.2.	Entonación . . . . .	76
3.2.3.	Acento léxico . . . . .	78
3.2.4.	La importancia de la entonación . . . . .	79
3.2.4.1.	La importancia de la entonación en general . . . . .	79
3.2.4.2.	La importancia de la entonación en la enseñanza de segundo idioma . . . . .	80
3.3.	El sistema propuesto . . . . .	83
3.3.1.	El sistema de evaluación de entonación . . . . .	84
3.3.1.1.	Pre-procesamiento . . . . .	84
3.3.1.2.	Extracción de F0 y post-procesamiento . . . . .	85
3.3.1.3.	Alineamiento DTW . . . . .	86
3.3.1.4.	Medida de similitud de F0 . . . . .	87
3.3.2.	El sistema de evaluación de acento léxico . . . . .	90
3.4.	Experimentos . . . . .	92
3.4.1.	Bases de datos . . . . .	92
3.4.1.1.	Base de datos para evaluación de entonación . . . . .	92
3.4.1.2.	Base de datos para evaluación de acento léxico . . . . .	93
3.4.1.3.	Configuración experimental . . . . .	94
3.4.1.4.	Puntuación basada en la correlación objetiva-subjetiva	95

3.4.2.	Experimentos para medir la exactitud del alineamiento DTW . . . . .	97
3.5.	Resultados y discusiones . . . . .	99
3.5.1.	Experimentos de alineamiento . . . . .	99
3.5.2.	Experimentos de entonación . . . . .	102
3.5.3.	Experimentos de acento léxico . . . . .	104
3.6.	Conclusiones . . . . .	107
<b>4.</b>	<b>Modelación prosódica para detección de emociones usando modelos</b>	
	<b>de referencia</b>	<b>109</b>
4.1.	Introducción . . . . .	110
4.2.	Antecedentes . . . . .	116
4.2.1.	Trabajo relacionado . . . . .	116
4.2.2.	Bases de datos emocionales . . . . .	119
4.2.3.	Extracción de F0 y post-procesamiento . . . . .	121
4.3.	Análisis de la prominencia emocional usando una única señal como re-	
	ferencia . . . . .	122
4.4.	Análisis de la prominencia emocional usando una familia de funciones .	126
4.4.1.	Extensión del enfoque propuesto para modelar la variabilidad	
	inter-locutor e intra-locutor en el contorno de F0 . . . . .	126
4.4.2.	Análisis discriminante . . . . .	131
4.4.3.	Bases léxico-dependiente versus léxico-independiente . . . . .	136
4.5.	Análisis y evaluación de la prominencia emocional a nivel de sub-oración	139

4.6. Conclusiones . . . . .	147
<b>5. Conclusiones</b>	<b>149</b>
<b>6. Referencias</b>	<b>155</b>

# Índice de figuras

2.1. Enfoque seguido en la tesis. . . . .	26
2.2. El aparato fonador humano. . . . .	28
2.3. Diagrama de bloques de la parametrización acústica de la señal de voz. . . . .	37
2.4. Segmentos sonoros (a) y sordos (b) de una señal de voz femenina. . . . .	38
2.5. Comparación de curvas punto a punto (a), y usando alineamiento temporal (b). . . . .	44
2.6. Representación “activación-evaluación”. El eje horizontal diferencia las emociones positivas y negativas, mientras que el eje vertical discrimina entre estados emocionales activos y pasivos. . . . .	63
2.7. Diagrama en bloques del enfoque estándar en detección de emociones. . . . .	64
3.1. Diagrama en bloques del sistema de evaluación de entonación para enseñanza de idiomas. . . . .	84
3.2. Diagrama en bloques del sistema de evaluación de acento léxico propuesto. . . . .	91

3.3.	Representación de la medida de error, $d$ , para el alineamiento DTW. El punto $(b_i^R, b_i^S)$ indica la intersección del límite $i$ dentro de las señales de referencia y test. Las distancias $d_R$ y $d_S$ son las distancias horizontales y verticales, respectivamente, entre los límites fonéticos y el alineamiento DTW. . . . .	99
3.4.	Correlación subjetiva-objetiva promedio en evaluación de entonación para diferentes micrófonos. Mic1 representa el micrófono de alta calidad, mientras que Mic2 y Mic3 corresponden a micrófonos para computador de escritorio de bajo costo. . . . .	101
3.5.	Curva ROC ( <i>false negative</i> versus <i>false positive</i> en evaluación de acento léxico. La medida de similitud fue calculada de acuerdo a la ecuación 3.11 y la decisión se estimó usando la ecuación 3.12. El valor $\alpha = 1$ indica que se utiliza sólo el contorno de F0 y $\alpha = 0$ indica que solamente la energía. . . . .	106
4.1.	Distribución de la medida de similitud con las emociones neutra (neutral), enojo (angry), felicidad (happy) y tristeza (sad) en la base de datos EMA. La medida de similitud corresponde a la correlación de Pearson entre los contornos de F0 neutros y emocionales. . . . .	127
4.2.	Marco general del método propuesto: (a) generación de modelos neutrales usando Functional PCA; y, (b) proyección de una señal de test en el espacio neutral. . . . .	128

4.3.	Reconstrucción de los contornos de F0 usando Functional PCA: (a) datos de entrenamiento para generar la base neutral usando funcional PCA; (b) reconstrucción de una señal de test neutra con las cinco primeras componentes principales; y, (c) reconstrucción de una elocución “feliz” (happy) usando las cinco primeras componentes principales. El error cuadrático medio entre el contorno de F0 original y reconstruido es igual a 0,45 y 0,32 para las señales neutra y feliz, respectivamente. . . . .	130
4.4.	Valor absoluto promedio de las proyecciones asociadas a cada componente principal obtenida con la base de datos EMA. . . . .	131
4.5.	(a) Ejemplo de la métrica subjetiva (punteada), derivada de la subjetiva (segmentada) y objetiva (continua). En este ejemplo, la correlación entre las métricas objetiva y subjetiva es igual a $\rho = 0,51$ . (b) Ejemplo de la métrica subjetiva (punteada), derivada de la subjetiva (segmentada) y objetiva (continua). En este ejemplo, la correlación entre las métricas objetiva y la derivada de la métrica subjetiva es igual a $\rho = 0,55$ . . . .	143

# Índice de Tablas

3.1. Escala subjetiva estricta (a) y no-estricta (b) para comparación de entonaciones. . . . .	97
3.2. Error de alineamiento DTW usando distintas características. La distancia local corresponde a la métrica Euclidiana. . . . .	100
3.3. Error de alineamiento DTW con y sin mismatch de locutor. . . . .	101
3.4. Error de alineamiento DTW con y sin mismatch de micrófono. . . . .	101
3.5. Correlación subjetiva-objetiva promedio en evaluación de entonación con distintas medidas de similitud. Las escalas estricta y no-estricta están definidas en la Tabla 3.1. . . . .	102
3.6. Correlación subjetiva-objetiva promedio en evaluación de entonación con distintas medidas de similitud. Se comparan las condiciones con y sin mismatch de locutor, usando la escala no-estricta definida en la Tabla 3.1.	103

3.7.	Correlación subjetiva-objetiva promedio en evaluación de entonación con distintas medidas de similitud. Se comparan las condiciones con y sin mismatch de pronunciación de segmentos, usando la escala no-estricta definida en la Tabla 3.1. . . . .	104
3.8.	Área bajo la curva ROC y equal error rate (EER) para evaluación de acento léxico para distintos valores de $\alpha$ usando correlación como medida de similitud. El $\alpha$ óptimo que minimiza el EER es 0,49. . . . .	106
4.1.	Descripción de las bases de datos. . . . .	123
4.2.	Análisis discriminante para las proyecciones obtenidas con funcional PCA a nivel de oración usando bases léxico dependientes con las bases de datos EMA y EMO-DB ( $Acc = Accuracy$ , $Pre = Precision$ , $Rec = Recall$ , $F = F\text{-score}$ ). Chance corresponde al número total de muestras emocionales dividido por el número total de señales. . . . .	134
4.3.	Desempeño del sistema de referencia con las bases de datos EMA y EMO-DB. Las características fueron extraídas a partir del contorno de F0 a nivel de oración ( $Acc = Accuracy$ , $Pre = Precision$ , $Rec = Recall$ , $F = F\text{-score}$ ). Chance corresponde al número total de muestras emocionales dividido por el número total de señales. . . . .	135

- 4.4. Test de hipótesis (proporciones) para determinar si las diferencias entre los clasificadores son estadísticamente significativas en la base de datos EMA. Los colores claros y oscuros representan significancia estadística fuerte ( $p\text{-value}<0.05$ ) y débil ( $p\text{-value}<0.1$ ), respectivamente ( $R$  = Sistema de Referencia,  $LD$  = modelos léxico-dependientes,  $LI$  = modelos léxico-independientes,  $Vf$  = segmentación basada en ventanas de tamaño fijo,  $Ch$  = Segmentación a nivel de chunk,  $Pa$  = segmentación a nivel de palabras) . . . . . 136
- 4.5. Análisis discriminante para las proyecciones obtenidas con funcional PCA a nivel de oración para bases léxico-independientes con las bases de datos EMA y EMO-DB ( $Acc$  = Accuracy,  $Pre$  = Precision,  $Rec$  = Recall,  $F$  = F-score). Chance corresponde al número total de muestras emocionales dividido por el número total de señales. . . . . 138
- 4.6. Exactitud (accuracy) para diferentes niveles de segmentación usando las bases de datos EMA y EMO-DB con modelos léxico-independientes. En el corpus EMO-DB, los resultados para los niveles Chunk y palabra no se entregan dado que la segmentación a nivel de fonemas no está disponible. 141

4.7. Inter-evaluator agreement (IEA) y correlaciones subjetiva-objetiva con la base de datos SEMAINE para diferentes tamaños de ventana (*IEA* = correlación entre la evaluación subjetiva de un sujeto y el promedio de los restantes evaluadores en la base de datos,  $\rho(S, O)$  = correlación entre el promedio de las evaluaciones subjetivas y la métrica objetiva,  $\rho(\frac{\Delta S}{\Delta t}, O)$  = correlación entre la derivada del promedio de las evaluaciones subjetivas y la métrica objetiva). . . . . 141

# Capítulo 1

## Introducción

### 1.1. Modelación prosódica

La idea de desarrollar máquinas capaces de realizar tareas intrínsecamente humanas como hablar, distinguir sonidos o percibir emociones ha estado presente durante años y ha sido materia de investigación de muchos científicos de las más diversas áreas. Este interés yace en la creciente necesidad de interfaces hombre-máquina más naturales y que sean cada vez más parecidas a las interacciones humanas. La ciencia ha permitido crear máquinas capaces de realizar labores automáticas de manera eficiente, con una precisión y velocidad elevadas, como por ejemplo cálculos aritméticos. Sin embargo, tareas como distinguir voces o reconocer rostros, que por cierto son bastante naturales para humanos, han significado verdaderos desafíos al momento de ser desarrollados artificialmente. En efecto, muchos de estos problemas aún no han podido ser resueltos

a cabalidad.

En este contexto, las tecnologías de voz juegan un rol importantísimo, ya que no tan sólo han impactado en industria de las telecomunicaciones y de los sistemas multimedia, sino que también otras áreas como telemática, juegos, automóviles, sistemas para personas con discapacidad, medicina y educación. El reconocimiento o la síntesis de voz, tecnologías que han sido desarrolladas desde fines de la década de los 70, han mostrado un fuerte crecimiento gracias a la masificación de los sistemas informáticos e Internet.

En los últimos años nuevos problemas han despertado el interés de los investigadores del área de procesamiento de voz. Entre ellos, el análisis y modelación prosódica (i.e. entonación, ritmo, acentuación y duración) es una disciplina que ha mostrado un crecimiento vertiginoso dada su importancia y variadas aplicaciones. Por ejemplo, características como la naturalidad de un sintetizador de voz o la calidad de un vocoder dependen directamente de la prosodia. Algunos autores han considerado este aspecto para mejorar el desempeño de sistemas de reconocimiento de voz y verificación de locutor cite [1, 2].

En el proceso de enseñanza de idioma la prosodia es un elemento fundamental, ya que provee al hablante de características esenciales en la comunicación como naturalidad, fluidez, intención y actitud. El uso correcto de la entonación o el acento son importantes cuando se adquiere un idioma. Por lo tanto, un sistema de enseñanza de idioma asistido por computador debe proveer un módulo mediante el cual los estu-

diantes puedan entrenar la percepción y producción de prosodia. En la literatura se ha estudiado el problema de evaluación de pronunciación, pero no de forma exhaustiva. Esta tesis aborda este problema mediante la propuesta de sistemas novedosos para evaluar automáticamente la entonación y el acento léxico en enseñanza de segundo idioma.

Otra aplicación interesante de la modelación y análisis de prosodia es el área de computación afectiva, disciplina relativamente nueva que ha ganado importancia en los últimos años debido a sus potenciales aplicaciones. Las emociones son fundamentales en la percepción, aprendizaje, comunicación, e incluso se relacionan con la toma de decisiones racionales e inteligentes y varias otras funciones cognitivas [3], y muchas veces son ignoradas en el diseño de interfaces hombre-máquina. Esto puede deberse a que las emociones son difíciles de medir. Además, la emoción está relacionada con el canal de comunicación implícito [4], el que lógicamente es mucho más difícil de abordar que el canal explícito. En esta tesis se propone una técnica para estimar la prominencia emocional en señales de voz usando modelos de referencia.

## **1.2. Entonación y acento en enseñanza de idiomas**

A pesar de la importancia de elementos como la entonación y el acento en la comunicación oral, el problema de enseñanza de idiomas asistido por computador (CALL, *computer aided language learning*) desde el punto de vista de la prosodia es un campo que ha sido muy poco explorado por investigadores del área. Además, en los métodos

existentes no está bien definida la separación entre la pronunciación de sonidos individuales y los aspectos prosódicos (entonación, ritmo, acento, fluidez, etc.) cuando se evalúa la pronunciación. Por tanto, este tema resulta ser complejo y de especial interés para el área de procesamiento de voz. La estrategia propuesta involucra la generación de un puntaje o nota que represente la calidad de la entonación de una elocución dada. Para la evaluación de acento, se utilizan técnicas de reconocimiento de patrones para determinar si el acento léxico de un estudiante es correcto o no.

Vale la pena mencionar el impacto tanto tecnológico como social de una aplicación de la enseñanza de idiomas, lo cual hace que sea de gran interés. Un sistema que permite evaluar automáticamente aspectos como la entonación y el acento, en conjunto con herramientas como Internet, permite masificar una aplicación novedosa que ciertamente puede significar un gran aporte a la educación, siendo de esta forma un apoyo tanto a los docentes como a los estudiantes para complementar la enseñanza presencial con un computador desde sus hogares, un laboratorio o desde cualquier lugar sin la supervisión constante de un profesor.

### **1.3. F0 y emociones**

Como resultado de los métodos usados para CALL, se proponen técnicas para la detección de emociones en señales acústicas, tarea que consiste en determinar si una elocución dada es emocional o no. Para ello, se evalúa la pertinencia del uso de modelos de referencia en voz emocional. Luego se presenta una estrategia basada en *functional*

*data analysis* para construir modelos de referencia robustos que incorporan variabilidad inter-locutor. Finalmente, se presenta un esquema para detectar emociones a nivel de sub-oración en tiempo real. El uso de modelos de referencia en detección de emociones no ha sido abordado previamente en la literatura.

Como se ha mencionado anteriormente, el problema de detección de prominencia emocional es de gran interés para el desarrollo de interfaces hombre-máquina que estén en sintonía con las necesidades de los usuarios. La propuesta presentada en este trabajo tanto para enseñanza de idiomas como para emociones corresponde a un trabajo original ya que no han sido publicadas anteriormente. En virtud de esto se han generado dos publicaciones en revistas ISI (ver anexo “Publicaciones del autor”).

## **1.4. Objetivos**

### **1.4.1. Objetivo general**

Proponer técnicas de modelación automática de prosodia en señales de voz.

### **1.4.2. Objetivos específicos**

- Investigar modelos de evaluación de entonación *top-down* y formular una estrategia para aplicar técnicas de reconocimiento de patrones al problema de evaluación de acento léxico.
- Proponer un método de modelación prosódica aplicado a detección de emociones

en señales acústicas usando la frecuencia fundamental, independiente del texto y del locutor. Además, proponer una estrategia que permita aplicar el sistema de detección emociones a nivel de sub-oración.

## **1.5. Estructura de la tesis**

Este trabajo ha sido estructurado de modo tal de presentar al lector pueda el marco general de este trabajo, entregando todos los conceptos, definiciones y antecedentes necesarios para comprender en detalle los temas abordados. De este modo, la tesis comienza con una introducción acerca de la prosodia en procesamiento de voz y sus aplicaciones en enseñanza de idiomas y reconocimiento de emociones. Luego se presenta una descripción detallada de la propuesta de esta tesis, considerando además resultados experimentales y comparaciones con técnicas en el estado del arte que se encuentran en la literatura. El trabajo se divide en cinco capítulos que son explicados brevemente a continuación.

El capítulo 2 muestra una revisión bibliográfica sobre tecnologías de voz, modelación y análisis de prosodia con el objetivo de introducir al lector a los problemas de evaluación de entonación y acento en enseñanza de idiomas y reconocimiento de emociones en señales acústicas. Este capítulo presenta definiciones conceptos relacionados con las características segmentales y suprasegmentales de la voz humana. Asimismo, se describen técnicas relacionadas con los temas abordados en esta tesis tales como métodos en el estado del arte, algoritmos de procesamiento de señales y metodologías

de evaluación.

El capítulo 3 se describe una técnica novedosa para evaluación de prosodia en el contexto de la enseñanza de idiomas. En particular, se desarrolla un sistema para la evaluación de entonación y acento léxico usando una estrategia *top-down*. El sistema compara directamente la señal del usuario con una señal pregrabada de referencia (modelo de referencia), entregando como resultado un puntaje o nota. En la comparación, se utilizan los parámetros acústicos de la prosodia y estrategias de reconocimiento de patrones. Los resultados indican que el desempeño de los sistemas propuestos permite utilizarlos en aplicaciones reales de CALL.

En el capítulo 4 extiende la idea de utilizar modelos de referencia presentada en el capítulo 3 al problema de detección de emociones en señales acústicas. Con el fin de abordar el problema de la variabilidad de locutores, se utiliza *functional data analysis* para generar modelos de referencia de la prosodia extraída de un conjunto de señales emocionalmente neutras que corresponden a una base de funciones ortogonales. Para determinar si una señal es emocional, se estiman las proyecciones de la señal de test sobre la base de funciones de referencia. Se realizaron diversos experimentos incluyendo bases de datos emocionales actuadas y naturales. Los resultados sugieren que el método puede ser empleado en aplicaciones reales.

Finalmente, en el capítulo 5 se presentan las conclusiones finales de los métodos propuestos en la tesis y se describen las principales direcciones de trabajo futuro.

## 1.6. Contribuciones de la tesis

En este trabajo se presenta un marco novedoso para la evaluación de prosodia basadas en modelos de referencia con aplicaciones en enseñanza de idiomas asistida por computador y detección de la modulación emocional en señales de voz. Respecto al método para educación, se pueden mencionar las siguientes contribuciones: una discusión exhaustiva detallada acerca del rol de la entonación en el aprendizaje de segundo idioma; un sistema independiente del texto y del idioma para evaluar la entonación; un sistema independiente del texto y del idioma para evaluar el acento léxico; y una evaluación de la robustez del alineamiento DTW (*dynamic time warping* o alineamiento temporal dinámico) respecto al locutor, pronunciación de segmentos y *mismatch* de micrófono para el sistema propuesto. Como resultado, se genera la siguiente publicación:

Arias, J.P., Yoma, N.B., Vivanco, H., (2010), “*Automatic intonation assessment for computer aided language learning*,” *Speech Communication* (Elsevier), Volume 52, Issue 3, March 2010, pp. 254-267.

Por su parte, en detección de emociones las principales contribuciones corresponden a: una técnica novedosa para la detección de emociones que permite modelar los contornos de F0 generados con voz neutra, independiente del texto y del locutor; un análisis detallado del uso de referencias neutras como un método para detectar las emociones en señales de voz; la generación de modelos de referencia de contorno F0 mediante *functional data analysis*; y un análisis de la unidad de segmentación más corta para

detección de emociones. Como resultado, se ha generado la siguiente publicación que está en proceso de *peer review*:

Arias, J.P., Busso, C., Yoma, N.B., (2012), “*Shape-based modeling of the fundamental frequency contour for emotion detection in speech,*” Submitted to Computer Speech and Language (Elsevier).

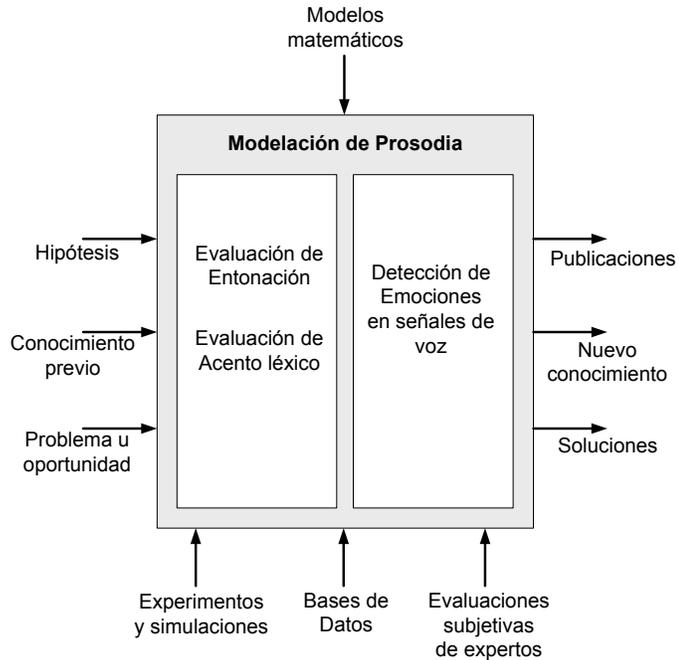
# Capítulo 2

## Revisión bibliográfica

### 2.1. Introducción

La Figura 2.1 describe el enfoque seguido en esta tesis, donde se muestra el trabajo desarrollado en función del objetivo general (modelación de la prosodia) y los dos objetivos específicos (evaluación de entonación y acento; y, detección de emociones en señales de voz). Los problemas u oportunidades, como las aplicaciones de la evaluación de prosodia en problemas de ingeniería, el conocimiento previo y las hipótesis de este trabajo permiten definir el dominio de la tesis. Por otra parte, la propuesta para resolver el problema se plasma en un modelo matemático. Los experimentos y simulaciones, junto con las bases de datos y las evaluaciones de expertos, permiten validar los modelos propuestos. Como resultado de este proceso, se obtienen soluciones y nuevo conocimiento, trabajo que se difunde a través publicaciones científicas que a su vez son

utilizadas por otros investigadores.



**Figura 2.1:** *Enfoque seguido en la tesis.*

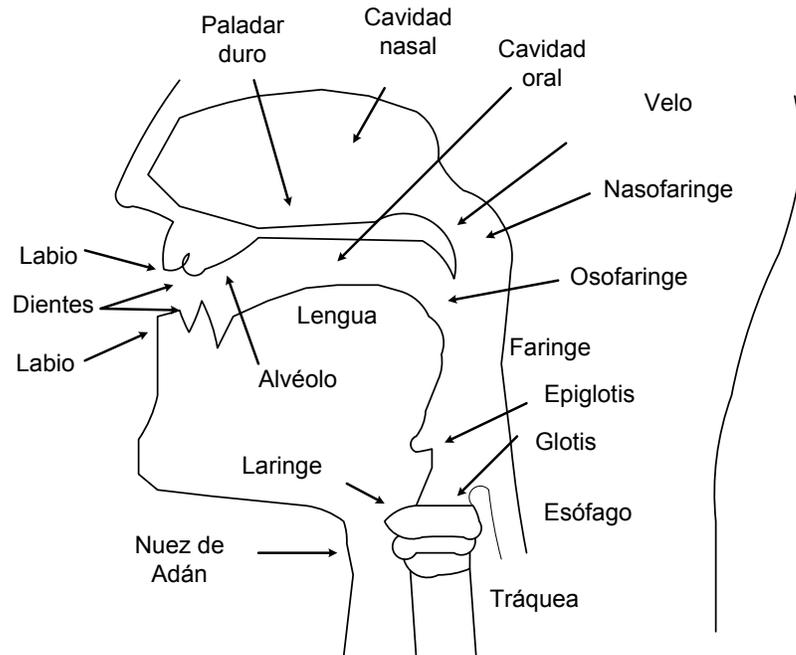
En este capítulo se introduce al lector a los problemas de evaluación de prosodia en educación y detección de emociones explicados en la secciones 1.2 y 1.3. Para ello, se presentan conceptos y técnicas de procesamiento de señales que son empleadas en esta tesis. Primero, se describen las características segmentales y suprasegmentales de la voz humana. Después, se explican en detalle algunos algoritmos usados en el procesamiento digital de la señal de voz y finalmente se realiza una revisión exhaustiva de la prosodia en enseñanza de idiomas y detección de emociones.

## 2.2. La voz humana

La voz se define como el sonido generado voluntariamente por el aparato fonador humano. Este aparato está compuesto 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 bronquios y la tráquea, pasando luego por las *cuerdas vocales* o *pliegues vocales*, ligamentos elásticos que se encuentran dentro de las paredes de la laringe. El espacio que existe entre las cuerdas vocales se denomina glotis. Este aire es modificado por las cavidades naso-buco-faríngeas, para ser finalmente moldeado por los articuladores. Las resonancias del tracto vocal producen modificaciones en el espectro de frecuencia del sonido que proviene de la laringe, determinando así los *formantes*.

La fuente de excitación del sistema de producción de voz humana se puede caracterizar de acuerdo a fonación, exhalación, fricación, compresión, vibración o una combinación de las anteriores [5]. Cuando los pliegues vocales están tensos y juntos pero no completamente cerrados, 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 español y el inglés, o consonantes como la  $/n/$ ;  $/m/$  y  $/l/$ . A mayor tensión en las cuerdas



**Figura 2.2:** *El aparato fonador humano.*

vocales, mayor es el valor de  $F_0$ .

La fricación se produce por constricciones del tracto vocal, generando un flujo de aire turbulento. Por su parte, la exhalación es producida por una abertura entre el cartílago aritenoides en la parte posterior de las cuerdas vocales. En estos dos casos, si las cuerdas vocales están lo suficientemente separadas permitiendo la libre expulsión del aire (excitación *noisy*), se producen sonidos *sordos* o *áfonos*, como la consonante  $/j/$  del español o el sonido *th* ( $\theta$ ) del inglés.

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 (impulso), por ejemplo el sonido  $/p/$ . La vibración se produce cuando el flujo de aire es parcialmente interrumpido en alguna parte del tracto vocal distinto a las cuerdas voca-

les (e.g. la lengua o los labios). Cabe mencionar que se pueden generar combinaciones de los tipos de excitación explicados anteriormente. Por ejemplo, al combinar fonación con fricación se producen sonidos como la /z/ del inglés, que se articula de manera similar a la /s/ del español pero con las cuerdas vocales tensas. Estos sonidos se caracterizan como fricativos *voiced* y en la literatura se habla de excitación semi-periódica.

Existen también sonidos que 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. De hecho, 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 y la capacidad pulmonar. En particular, la frecuencia fundamental que puede generar un locutor depende directamente del ancho y largo de sus cuerdas vocales, parámetros que están directamente relacionados con el género y la edad. De hecho, el rango de variación de F0 para las mujeres es mayor que en los hombres adultos.

### **2.3. Características segmentales**

En lingüística, se definen como características segmentales o segmentos a los sonidos individuales de la voz, relacionados con el lugar y la forma de articulación y están

relacionados en su mayoría con la región supralaríngea del aparato fonador humano [6]. A su vez, el lugar y la forma de articulación dependen de la posición de los labios, la altura de la lengua, los movimientos del velo del paladar y de la mandíbula. La única excepción es el aspecto *voice* (i.e. si un sonido es sonoro o áfono). Se considera además que la aspiración y la glotalización, sonidos presentes en lenguas como el hindi, también son propiedades de los segmentos pese a que se producen a nivel de laringe.

El habla continua es el resultado de la concatenación de varios segmentos diferentes. Cuando un locutor pronuncia una secuencia de sonidos se presenta un grado de acomodación entre segmentos adyacentes, fenómeno llamado coarticulación [6]. Por ejemplo, en la palabra “música” la posición de los labios durante el sonido */m/* anticipa de alguna manera el segmento */u/* (coarticulación anticipatoria). Otro ejemplo se da en la palabra *tenth* en inglés, donde el sonido */n/*, que normalmente se pronuncia posicionando la lengua en los alvéolos, se pronuncia con una articulación dental.

Desde el punto de vista de procesamiento de voz, los segmentos se analizan mediante análisis espectral de corta duración (*short-term analysis*). Para ello, las señales acústicas se dividen en cuadros de duración relativamente cortos (usualmente entre 20 y 25 milisegundos) y cada ventana se representa en frecuencia mediante el análisis de Fourier (la sección 2.5.1 se entrega mayores detalles de este proceso). De esta manera un conjunto de cuadros adyacentes puede representar un segmento. Los sistemas de reconocimiento automático de voz utilizan como unidad acústico-fonética el trifonema, el que se compone de una unidad central más un segmento que lo antecede y otro

que lo sucede, modelando de esta manera la coarticulación. La técnica más usada en reconocimiento de voz consiste en representar las palabras, frases u oraciones mediante secuencias de trifonemas usando modelos ocultos de Markov (HMM, *hidden Markov models*) con una topología *left-to-right* [7].

## 2.4. Características suprasegmentales

Las características suprasegmentales o prosódicas aparecen a en unidades fonéticas mayores a los segmentos individuales. Algunos autores definen prosodia como aquellas características cuyo dominio se extiende más allá de un segmento (por ejemplo una sílaba o una elocución) [8]. Otros investigadores también hablan de una variación o modificación secundaria de los segmentos. Desde el punto de vista físico, las características suprasegmentales ocurren principalmente en la laringe y la región subglotal. La entonación y el tono son controlados por los músculos de la laringe, mientras que la intensidad dependen del flujo de aire emanado de los pulmones, que a su vez es controlado por los músculos respiratorios [6]. A continuación se definen los aspectos prosódicos más importantes y sus funciones en el habla.

### 2.4.1. Entonación

De acuerdo a Botinis *et al.*, la entonación se define como la combinación de estructuras tonales dentro de unidades estructurales más grandes asociadas con el parámetro acústico de la voz denominado frecuencia fundamental o F0 y sus variaciones distinti-

vas en el proceso del habla [9]. Por su parte, F0 se define como el número de ciclos por segundo en que las cuerdas vocales completan un ciclo de vibración. En consecuencia, la producción de la entonación es regulada por las fuerzas musculares de la laringe que controlan la tensión de las cuerdas vocales, en conjunto con las fuerzas aerodinámicas del sistema respiratorio. La percepción de la entonación se denomina *pitch*.

La entonación posee varias funciones lingüísticas que vale la pena mencionar [10, 11, 12, 13, 14]:

1. Función *actitudinal*. Hace posible expresar emociones y actitudes en el habla, por ejemplo, si un saludo es rutinario o entusiasta.
2. Función de *prominencia*. La entonación tiene un rol significativo cuando se asigna prominencia a las sílabas que deben ser reconocidas como acentuadas. 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” [15]. Otros lo han denominado *acento de insistencia*, *acento enfático* [16] y *acento expresivo* [17], siendo este último el más ampliamente aceptado.
3. Función *gramatical*. Esta función permite al interlocutor reconocer fácilmente la estructura gramatical y sintáctica de una oración dada, por ejemplo, para distinguir entre una sentencia afirmativa de una interrogativa.

4. Función de *discurso*. Si se considera el habla desde una perspectiva amplia, se puede observar que la entonación puede sugerir al interlocutor qué puede ser tomado como nueva información nueva y qué es considerado como información dada. En una conversación, puede proveer una indicación en relación al tipo de respuesta esperada. Por ejemplo: considerar la sentencia en inglés “*I sent the book to John on Tuesday*”. Si se hace énfasis en palabra “book”, el locutor estaría indicando que el libro enviado es relevante, y no la persona ni el día. Por el contrario, si el hablante hace énfasis en la palabra “Tuesday”, el interlocutor entenderá que es más relevante el día en que fue ejecutada la acción.
  
5. Función *léxico-semántica*. En las denominadas *lenguas tonales*, la variación de la frecuencia fundamental puede cambiar completamente el significado de una palabra. Por ejemplo, en el idioma chino mandarín, la secuencia de sonidos /ma/ puede tomar los siguientes significados modificando solamente su melodía: mamá, cáñamo, caballo y regañar. Muchos autores denominan a esta característica *tono* y la consideran distinta de la entonación. A pesar de que algunos investigadores hablan indistintamente de entonación y tono, en esta tesis son considerados conceptos diferentes.
  
6. Función de *naturalidad*. Los locutores nativos pueden reconocer fácilmente si una sentencia dada 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 (i.e.

pronunciación de segmentos); la entonación; y, el ritmo. Un locutor no nativo pero competente puede desviar la atención de su interlocutor, si independiente de su entonación (en conjunto con el ritmo) no es el mismo usado por un hablante nativo en las mismas circunstancias, ya que su habla se oiría poco natural.

### **2.4.2. Acento**

Algunos autores evitan el uso de la palabra “acento” ya que, como se menciona en [13], se emplea en fonética y lingüística en formas diversas y poco claras. A veces se emplea como un equivalente de intensidad o de prominencia. En esta propuesta de tesis se utiliza la definición presentada en [14]: “el acento está dado por una combinación de intensidad (energía), pitch y duración”.

En una palabra en inglés como “university”, la sílaba “ver” recibe el acento primario, mientras que en “u” recae el acento secundario. Las otras sílabas “ni”, “si” y “ty” se consideran no acentuadas. La presencia de sílabas que reciben acento primario o secundario es de interés en idiomas como el inglés, ya que tienden a pronunciarse completamente, mientras que en las sílabas no acentuadas suele presentarse el fenómeno conocido como reducción vocálica, que consiste en cambios en las características acústicas de las vocales.

Es de especial interés el acento que puede cambiar el significado de una palabra de acuerdo a la sílaba que lo contenga, denominado acento léxico. Por ejemplo, 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”; “límite”, “limite”, “limité”; “célebre”, “celebre”, “celebré”.

### **2.4.3. Duración**

En lingüística, la duración corresponde al tiempo que permanece un determinado segmento [14]. En algunos idiomas, la duración tiene un significado léxico-semántico, adquiriendo por tanto relevancia. Por ejemplo, en italiano la palabras “vile” (villano) y “ville” (villas). En inglés y en español la duración tiene efectos significativos, pero no léxico-semánticos. Por ejemplo, la palabra “no” pronunciada con una “n” más larga probablemente será interpretada como vacilante. Es interesante notar por ejemplo las diferencias del sonido “ch” en el español chileno. En la pronunciación estándar, este sonido tiene una duración promedio de 0,07 segundos, mientras que en la pronunciación de individuos pertenecientes a clases sociales bajas es de 0,14 segundos [18].

### **2.4.4. Otras características prosódicas**

Además de las características suprasegmentales explicadas anteriormente, se tiene el ritmo, que corresponde a un patrón perceptual producido por la interacción en el tiempo de la prominencia relativa de sílabas acentuadas y no acentuadas [6]. Esta característica guarda estrecha relación con la entonación. Se tiene también la tasa de sílabas por segundo o *rate*. Finalmente, la continuidad representa la cantidad de pausas usadas por un locutor mientras habla.

## 2.5. Algoritmos y técnicas utilizadas en procesamiento de voz

### 2.5.1. Parametrización acústica

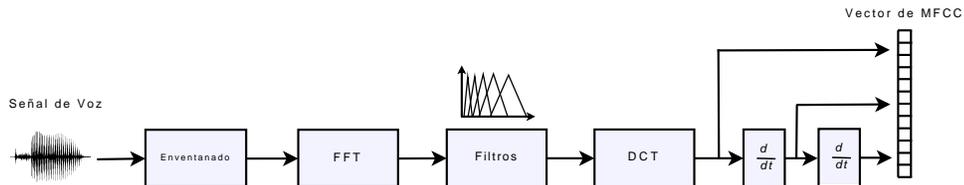
La conversión análogo-digital es la primera etapa en el pre-procesamiento, ya que los computadores requieren una representación discreta de la señal. En aplicaciones con interfaces hombre-máquina, esta conversión es efectuada por una tarjeta de sonido. Luego, se aplica un detector de inicio y fin de señal útil (denominado también *endpoint detector*), el cual se encarga de eliminar los segmentos de silencio iniciales y finales dado que éstos no contienen información acústica relevante. Después, se aplica un proceso de inventanado que consiste en dividir la secuencia que representa la voz en unidades llamadas cuadros o *frames*. Para ello, se utiliza la técnica de inventanado de *Hamming* [19] con el fin de obtener una mejor representación de la señal en el dominio de la frecuencia.

La etapa siguiente es el análisis espectral en cada *frame*, para lo cual se utiliza la transformada rápida de *Fourier* (FFT, *Fast Fourier Transform*). El oído humano no es capaz de percibir frecuencias puntuales sino que distingue intervalos, razón por lo cual 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 2.1, para una

determinada frecuencia  $f$  medida en Hertz:

$$Mel(f) = 2595 \log_{10} \left( 1 + \frac{f}{700} \right). \quad (2.1)$$

Los filtros corresponden a funciones triangulares con ganancia unitaria en la frecuencia central, 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 que 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 en bloques del proceso de extracción de parámetros acústicos.

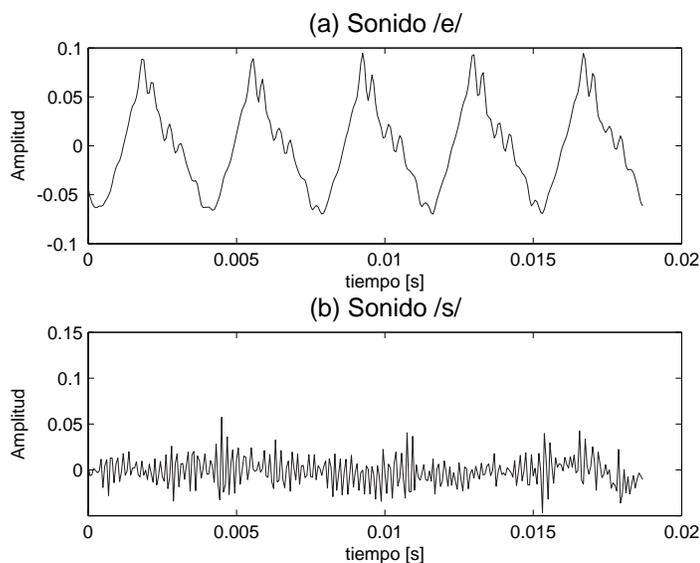


**Figura 2.3:** Diagrama de bloques de la parametrización acústica de la señal de voz.

### 2.5.2. Extracción de F0

La frecuencia fundamental o F0 corresponde a la frecuencia más baja de la descomposición armónica de una señal dada. Cuando se hace referencia específicamente a señales de voz, surgen varios problemas con la definición anterior debido a la no esta-

cionariedad y a la presencia de segmentos que no son periódicos. En este contexto, y de acuerdo a lo explicado en 2.2, se habla de sonidos sonoros cuando hay presencia de periodicidad en un determinado segmento de voz, y de sonidos sordos o áfonos cuando no la hay. Por ejemplo, en el español el sonido vocálico /e/ (Figura 2.4 (a)), o el sonido nasal /m/ se consideran sonoros, mientras que el sonido fricativo /s/ (Figura 2.4 (b)) es clasificado comúnmente como áfono. La estimación de la frecuencia fundamental consiste en decidir si un frame es sonoro o sordo (decisión voiced/unvoiced o V/UV), y si lo primero se cumple, encontrar el valor de F0.



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

Como se explica anteriormente, la determinación de la frecuencia fundamental juega un rol importante en las características suprasegmentales como el acento y la entonación [20]. Además, es de suma importancia en otras áreas como la síntesis de voz, codificación de voz, reconocimiento automático de voz de lenguas tonales como el chino

mandarín o el vietnamita, *speech enhancement*, verificación de locutor, etc. No obstante lo anterior, la correcta estimación de F0 no está exenta de dificultades. En primer lugar, es muy difícil evaluar un algoritmo ya que no se cuenta con una medida objetiva conocida. Además, el tracto vocal puede enfatizar otros armónicos. Es muy común que los armónicos o sub-armónicos de F0 sean erróneamente detectados como tal, lo cual da origen a los fenómenos denominados *doubling* y *halving*, respectivamente. Por último, el ruido proveniente de la respiración del locutor puede provocar que los segmentos sonoros parezcan áfonos. Además, cuando se utilizan filtros de banda muy angosta, los segmentos sordos pueden presentar una periodicidad aparente. En virtud de lo anterior, existen muchos trabajos que han abordado el problema de estimación de F0, entre los cuales se pueden encontrar métodos en el dominio del tiempo y de la frecuencia.

### Métodos en el dominio del tiempo.

Una técnica comúnmente utilizada es la estimación mediante la función de autocorrelación propuesta por Rabiner *et al.* [21]. Si  $x(n)$  es un *frame* proveniente de una señal de voz, entonces su función de autocorrelación,  $r_{xx}$ , está definida por:

$$r_{xx}(\tau) = \sum_{t=1}^{N_m} x(t) \cdot x(t - \tau) \quad (2.2)$$

donde  $N_m$  es el número de muestras del frame y  $\tau$  es el desplazamiento temporal. Si se considera que la señal es periódica con período  $P$ , entonces la función  $r_{xx}$  tendrá máximos locales en  $r_{xx}(\tau + iP)$ ,  $\forall i \in \{0, \dots, N_m\}$ . En este caso, el segundo máximo local

corresponderá al período fundamental  $T0$ , y  $r_{xx}(T0)$  entrega una medida de periodicidad de la señal. En la literatura existen muchas variantes de este método [22, 23], siendo ampliamente utilizada la autocorrelación normalizada que tiene la ventaja de ser independiente de la energía de la señal:

$$r_{xx}(\tau) = \frac{\sum_{t=1}^{N_m} x(\tau) \cdot x(n - \tau)}{\sqrt{\sum_{t=1}^{N_m} x^2(\tau) \cdot \sum_{t=1}^{N_m} x^2(\tau - N_m)}}. \quad (2.3)$$

Por otra parte, se tiene el método YIN, el que fue introducido por de Cheveigné *et al.* [24], 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  $x(n + \nu)$ , donde  $\nu$  es un desplazamiento dado, en vez de maximizar su producto:

$$d_t(\nu) = \sum_{n=1}^W (x(n) - x(n - \nu))^2. \quad (2.4)$$

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:

$$d'_t(\nu) = \begin{cases} 1 & \text{si } \nu = 0 \\ \frac{d_t(\nu)}{\frac{1}{\nu} \sum_{j=1}^{\nu} d_t(j)} & \text{en otro caso} \end{cases} \quad (2.5)$$

Finalmente, se añaden etapas de interpolación y de eliminación de errores de octava.

## Métodos en el dominio de la frecuencia.

Dentro de esta familia se tiene el análisis de cepstrum, el cual 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 periodicidad en el espectro de la señal a través de máximos locales en el dominio cepstral.

Otro método en el dominio de la frecuencia es máxima verosimilitud [25]. Este algoritmo busca entre un conjunto de posibles espectros  $\tilde{X}$ , definidos como un tren de impulsos de cierta frecuencia  $\omega$ , y escoge aquel que mejor calza con la forma del espectro de entrada  $X$ . El error que se comete al aproximar  $X$  por el espectro ideal,  $E(\omega)$ , se define como:

$$E(\omega) = \|X(\omega) - \tilde{X}(\omega)\| \quad (2.6)$$

$$\Rightarrow E(\omega) = \|X(\omega)\|^2 + \|\tilde{X}(\omega)\|^2 + 2X(\omega)\tilde{X}(\omega)^T \quad (2.7)$$

$$(2.8)$$

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

$$\hat{X} = \min_{\omega} \{E(\omega)\} = \min_{\omega} \{X(\omega)\tilde{X}(\omega)^T\} \quad (2.9)$$

Basado en este mismo principio se han propuesto técnicas de análisis por síntesis [26], donde se compara la señal original con una versión sintetizada que depende del valor de la frecuencia fundamental  $F_0$ , generando una función objetivo que es optimizada. Además, el valor de la función de error  $E(\omega)$  puede ser utilizada para tomar la decisión sonoro/sordo.

### Otras técnicas

En la literatura existen muchos métodos y algoritmos que no pertenecen a las familias mencionadas anteriormente. Entre éstas se tienen: redes neuronales [27], lógica difusa [28], transformada de Hilbert-Huang [29] y transformada wavelet [30]. Existen además técnicas que se basan en la hipótesis que el pitch no puede cambiar abruptamente entre un frame y otro. Luego, es posible utilizar los frames adyacentes para incorporar información temporal a la detección de pitch. De esta forma se emplean métodos de programación dinámica (DP, *dynamic programming*) [31, 32, 33]. Esto se denomina *pitch tracking*, y tiene la ventaja de entregar una detección más exacta, muy útil para aplicaciones donde el procesamiento de la señal de voz se hace *offline*, como en un sistema de enseñanza de idiomas o de detección de emociones.

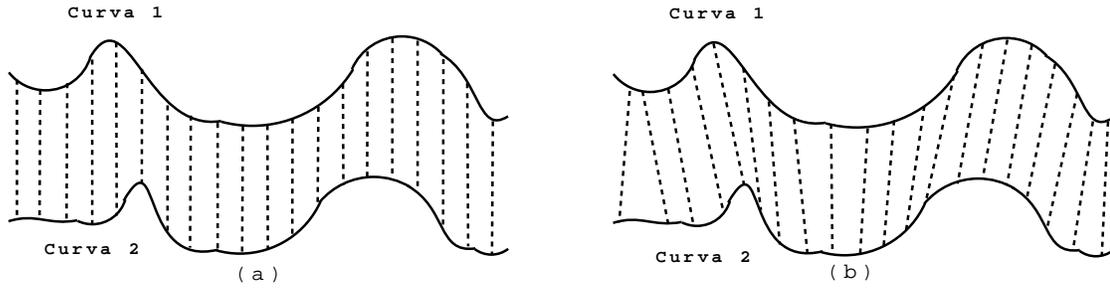
Una vez que ya se ha calculado el  $F_0$  con algún método de los revisados anteriormente, es posible utilizar algunas técnicas con el fin de eliminar algunos errores en la estimación (e.g. ruido, *halving*, *doubling* o problemas en la decisión V/UV). El uso de filtros medianos ha sido un método tradicional de post-procesamiento [34] el cual elimina valores erróneos y además produce un efecto suavizado o *smoothing*. Además,

existen técnicas más robustas como aplicación de métodos estadísticos [35], heurísticos [36] y de análisis en el dominio de la frecuencia [37].

### 2.5.3. El algoritmo DTW

En este trabajo se emplea el concepto de modelos de referencia, el que consiste en comparar una elocución de test con una referencia *frame a frame*. En esta comparación es esencial que ambas señales estén alineadas en el tiempo. Por tanto, para validar e implementar los sistemas propuestos para enseñanza de idiomas y para detección de emociones en señales acústicas es fundamental el algoritmo DTW que se describe a continuación.

Sean  $X$  e  $Y$  dos secuencias de vectores de distinta duración que se desean comparar con la finalidad de estimar una métrica de distancia o similitud. Lógicamente, las secuencias no son directamente comparables mediante técnicas estándar como la distancia euclidiana o Mahalanobis debido a que  $X$  e  $Y$  no están temporalmente alineadas. Por ejemplo, si se desea comparar dos señales de voz, no es válido usar una métrica que estime una diferencia o similitud *frame a frame*, ya que una elocución dada no puede ser pronunciada siempre a una misma velocidad, ni tampoco puede reproducirse de igual forma por dos locutores distintos. En la Figura 2.5 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 2.5:** 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 secuencias de observaciones que eventualmente pueden tener distinto largo. La idea es modificar de forma no lineal la dimensión temporal de ambas secuencias y mapearlas a un único conjunto de índices, de tal forma que la distancia calculada componente a componente sea mínima. Así, las zonas donde  $X$  es similar a  $Y$  quedan “alineadas”. Supóngase que secuencias  $X$  y  $Y$  pueden ser representadas por  $(x_1, x_2, \dots, x_{T_x})$  y  $(y_1, y_2, \dots, y_{T_y})$ . Además, se definen dos funciones de alineamiento  $\phi_x$  y  $\phi_y$ , las cuales transforman los índices de las secuencias de vectores a un eje  $k$  normalizado en el tiempo:

$$i_x = \phi_x(k), k = 1, 2, \dots, T$$

$$i_y = \phi_y(k), k = 1, 2, \dots, T$$

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

mediante la siguiente expresión:

$$d_\phi(X, Y) = \frac{\sum_{k=1}^T d(\phi_x(k), \phi_y(k))w(k)}{\sum_{k=1}^T w(k)} \quad (2.10)$$

En esta expresión,  $d(x, y)$  es una medida de distancia y  $w(k)$  corresponde a un vector de pesos. De acuerdo con la ecuación 2.10, el valor de la distancia  $d_\phi$  depende exclusivamente de las funciones de alineamiento que se escojan. Luego, el camino óptimo está dado por la siguiente ecuación (sujeta a ciertas restricciones):

$$d(X, Y) = \min_{\phi} d_\phi(X, Y). \quad (2.11)$$

Dado que el número de caminos posibles es muy grande es necesario imponer ciertas restricciones en la búsqueda [38]:

1. *Monotonía*, es decir  $\phi_x(k) \leq \phi_x(k+1)$  y  $\phi_y(k) \leq \phi_y(k+1)$ , para todo  $k$ .
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, consiste en imponer los puntos extremos del camino:  $\phi_x(1) = 1$ ;  $\phi_y(1) = 1$ ;  $\phi_x(T) = T_x$  y  $\phi_y(T) = T_y$ . 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 paralelogramo de Itakura.

5. *Pendiente.* Si se tiene una función de alineamiento  $\phi$  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  $k$ , entonces una parte de  $X$  se alinea con un segmento relativamente grande de  $Y$ . Para evitar este problema se imponen condiciones sobre la primera derivada de  $\phi$  permitiendo una cantidad limitada de movimientos horizontales y verticales.

La ecuación 2.10 es una expresión racional, por tanto, resolver el problema de optimización que plantea la ecuación 2.11 es extremadamente complejo. Si se define  $W_\phi$  como:

$$W_\phi = \sum_{k=1}^T w(k) \quad (2.12)$$

entonces, el problema planteado por la ecuación 2.11 se transforma en:

$$d(X, Y) = \frac{1}{W_\phi} \min_{\phi} \sum_{k=0}^T d(\phi_x(k), \phi_y(k)) w(k). \quad (2.13)$$

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. La primera es la *forma simétrica*, donde los pesos son escogidos como  $w(k) = (\phi_x(k) - \phi_x(k-1) + (\phi_y(k) - \phi_y(k-1)))$ . Así,  $W_\phi = T_x + T_y$ . Por otro lado, la *forma asimétrica* 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  $W_\phi = T_y$ .

Si  $D(i_x, i_y)$  se define como el *valor óptimo acumulado*, cuyo valor inicial está dado por  $D(1, 1) = 2d(1, 1)$  (forma simétrica), entonces  $D(i_x, i_y)$  se calculan de acuerdo a la siguiente recursión:

$$D(i_x, i_y) = \min \left\{ \begin{array}{l} D(i_x, i_y - 1) + d(i_x, i_y) \\ D(i_x - 1, i_y - 1) + 2d(i_x, i_y) \\ D(i_x - 1, i_y) + d(i_x, i_y) \end{array} \right\}. \quad (2.14)$$

Si se utiliza la forma asimétrica para los pesos, entonces la condición inicial es  $D(1, 1) = d(1, 1)$  y la ecuación es:

$$D(i_x, i_y) = \min \left\{ \begin{array}{l} D(i_x, i_y - 1) \\ D(i_x - 1, i_y - 1) + 2d(i_x, i_y) \\ D(i_x - 1, i_y) + d(i_x, i_y) \end{array} \right\}. \quad (2.15)$$

Las ecuaciones 2.14 y 2.15 no incluyen las restricciones de pendiente, luego corresponden a las de tipo  $P = 0$ . Para imponer condiciones de tipo  $P = 1$ , se puede establecer para la forma simétrica una ecuación recursiva de dos pasos:

$$D(i_x, i_y) = \min \left\{ \begin{array}{l} D(i_x - 1, i_y - 2) + 2d(i_x, i_y - 1) + d(i_x, i_y) \\ D(i_x - 1, i_y - 1) + 2d(i_x, i_y) \\ D(i_x - 2, i_y - 1) + 2d(i_x - 1, i_y) + d(i_x, i_y) \end{array} \right\}. \quad (2.16)$$

El algoritmo DTW ha sido ampliamente usado en reconocimiento automático de voz (ASR, *Automatic Speech Recognition*), tanto para palabras aisladas [38] como para habla continua [39]. Es considerado como el primer método que condujo a resultados aceptables en esta área. Sin embargo, ha sido desplazado por los modelos ocultos de Markov (HMM) ya que entrega mayor robustez y escalabilidad. No obstante 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, entre los cuales se pueden mencionar minería de datos [40], procesamiento de imágenes [41], bioinformática [42] y medicina [43].

#### **2.5.4. *Functional Data Analysis (FDA)***

Uno de los objetivos de esta tesis es modelar la frecuencia fundamental F0 con el fin de capturar la información emocional contenida en ella. Para lograr esto es necesario contar con un marco que permita representar y comparar curvas. En el contexto de este trabajo, donde se pretende capturar la forma de F0 y generar modelos de referencia usando un conjunto de curvas, *functional data analysis* resulta ser una alternativa interesante.

FDA es un conjunto de técnicas que permiten representar la estructura de las señales como funciones usando métodos estadísticos [44]. Dado un conjunto de datos, el objetivo principal de FDA es encontrar una representación que permita analizar sus ca-

racterísticas, patrones y variaciones. Los datos pueden ser representados mediante una función continua y suave,  $x(t)$ , la que a su vez se genera mediante una combinación lineal de funciones base  $\phi_k(t)$ :

$$x(t) = \sum_{k=1}^K c_k \phi_k(t). \quad (2.17)$$

donde  $K$  representa la dimensión de la expansión y  $c_k$  corresponde a la proyección en la  $k$ -ésima función base. Los datos funcionales se observan como una secuencia discreta  $(t_j, y_j)$ ,  $j \in \{1, \dots, n\}$ , donde  $y_j$  corresponde al valor muestreado de la función  $x(t)$  en  $t_j$ . La secuencia no está necesariamente muestreada a intervalos de tiempo regulares y además puede estar corrupta por ruido:

$$y_j = x(t_j) + \epsilon_j. \quad (2.18)$$

En FDA, el proceso de ajustar funciones a un conjunto de datos se denomina suavizado o *smoothing*. Dada una secuencia de observaciones discreta  $y_j$  y una base de funciones  $\{\phi_1, \dots, \phi_K\}$ , el suavizado intenta encontrar los coeficientes óptimos  $c_k$  minimizando el error  $\epsilon_k$ . Un parámetro de penalización de *roughness* es incorporado en la optimización para asegurar una representación suave. Los coeficientes  $c_k$  óptimos,  $\hat{c}_k$ , son estimados mediante la siguiente expresión:

$$\hat{c}_k = \underset{c_k}{\operatorname{argmin}} \sum_{j=1}^n [y_j - x(t_j)]^2 + \lambda \int [D^m x(s)]^2 ds \quad (2.19)$$

donde  $\lambda$  es un parámetro de suavizado (o *smoothing*) y  $D^m$  corresponde la  $m$ -ésima derivada [44]. Normalmente  $m \geq 2$  y representa la curvatura de los datos funcionales.

Los parámetros  $\phi_k(t)$  y  $K$  se deben escoger apropiadamente de acuerdo a las características de los datos. Dentro de las bases de funciones posibles se tienen *B-spline*, polinomial, wavelet, exponencial, Fourier y otras. Respecto a la elección del parámetro  $K$ , se debe tener en consideración que existe un *trade-off* entre el bias y la varianza. Si el número de funciones base  $K$  es grande, entonces la cantidad  $x(t) - E[\hat{x}(t)]$  tiende a cero, mientras que  $E[\{\hat{x}(t) - E[\hat{x}(t)]\}^2]$  disminuye para valores de  $K$  demasiado pequeños.

FDA ofrece varias ventajas cuando se compara con técnicas convencionales que representan datos como un conjunto de puntos aislados. La estadística descriptiva (e.g. media, covarianza y correlación) puede ser aplicada a los datos funcionales. Otra ventaja es que permiten representar las derivadas de los datos sin los problemas asociados al muestreo. Estas propiedades podrían ser especialmente útiles para modelar y analizar los contornos de F0. Además, poderosas herramientas para el análisis de datos tales como el análisis de componentes principales (PCA) pueden ser usados en el marco de FDA. En el contexto de esta tesis, vale la pena describir esta técnica denominada *functional PCA*.

El PCA tradicional es un método que convierte un conjunto de observaciones correlacionadas en variables no correlacionadas, llamadas componentes principales (PC, *principal components*), mediante el uso de transformaciones ortogonales. *Functional*

PCA es una técnica que extiende este concepto al dominio de las funciones [44]. Dado un conjunto de  $V$  funciones,  $x_v(t)$ , los *score* de componentes principales,  $f_{u,v}$ , están dados por:

$$f_{u,v} = \int \xi_u(t)x_v(t)dt \quad (2.20)$$

donde  $\xi_u(t)$  corresponde a una base ortonormal de funciones denominadas funciones de componentes principales o funciones PC (*principal components*) que representan la variabilidad de  $x_v(t)$ . La base de funciones es determinada de acuerdo al siguiente procedimiento:

1. Encontrar la primera función PC  $x_{i1}(t)$  mediante la maximización de  $\sum_v f_{1,v}^2$ , sujeto a la restricción:

$$\int \xi_1(t)^2 dt = 1. \quad (2.21)$$

2. Las funciones posteriores  $\xi_u(t)$  se obtienen mediante la maximización de  $\sum_v f_{u,v}^2$  sujeta a la restricción  $\int x_{iu}(t)^2 dt = 1$  y las  $m - 1$  restricciones adicionales:

$$\int \xi_u(t)\xi_m(t)dt = 0 \quad \forall m < u. \quad (2.22)$$

Finalmente, las funciones  $x_v(t)$  se puede aproximar mediante el uso de los primeros

$U$  componentes principales:

$$\hat{x}_v(t) = \sum_{u=1}^U f_{u,v} \xi_u(t). \quad (2.23)$$

La motivación del paso 1 es que la maximización del promedio cuadrático es identificar la dirección de mayor variación en las variables. La restricción asociada es fundamental para que el problema quede bien definido, ya que sin ésta la combinación de variables se puede crecer sin control. En el paso 2 (y en las iteraciones subsecuentes) también se identifican los modos de mayor variación, pero asegurando la ortogonalidad respecto a aquellas direcciones identificadas previamente.

FDA provee un marco interesante para modelar y analizar parámetros de las señales de voz y en especial la prosodia. Por ejemplo, Gubian *et al.* [45] utiliza FDA para analizar el fenómeno de reducción vocálica en el idioma francés. Para ello, analiza las transiciones dinámicas de la energía. Zellers *et al.* realizan un estudio de la entonación usando FDA [46]. Cheng *et al.* analizan el mecanismo de contracción tonal que ocurre en el mandarín de Taiwán, mediante el análisis de las trayectorias de F0 y su velocidad usando [47]. Como se puede apreciar, estos trabajos utilizan FDA como un marco para realizar un análisis descriptivo de los rasgos prosódicos. En el capítulo 4 de este trabajo se propone la aplicación de FDA como una herramienta para la generación de modelos neutrales y proyecciones para representar los contornos de F0 desde el punto de vista de reconocimiento de patrones.

## 2.6. Enseñanza de idiomas 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 Internet son utilizados para presentar los contenidos en forma interactiva. CALL ofrece enormes ventajas a los estudiantes como una herramienta complementaria a la instrucción presencial. La interactividad entregada por el software educativo hace más efectivo el proceso de aprendizaje ya que incrementa la motivación y permite desarrollar actividades y ejercicios sin necesidad de la supervisión permanente de un profesor. Los estudiantes que recién comienzan el proceso de adquisición de una lengua extranjera sienten temor e incomodidad al hablar en otro idioma, por lo que interactuar con un computador puede ayudar significativamente a aquellos más retraídos.

En los últimos años, CALL ha experimentado grandes cambios gracias al desarrollo de la ciencia y las tecnologías multimedia (sonido, animaciones, imágenes y texto) e Internet, así como también el avance de disciplinas como el procesamiento de señales, el reconocimiento automático de voz (ASR, *automatic speech recognition*) y la síntesis de voz (TTS, *text-to-speech*). En la actualidad, CALL se centra en la interacción social. En este contexto, 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 (HMI, *human-machine interface*) inteligentes.

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 sistemas utilizan un ASR dependiente del locutor con un vocabulario reducido. Este tipo de ejercicios han sido aplicados a la enseñanza de primer y segundo idioma [48]. Esta tecnología se caracteriza por su simplicidad y robustez.

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 alumnos mediante estímulos gráficos o simplemente a través de una pregunta directa. Existen 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, quien a su vez debe generar la secuencia de palabras más apropiada. La tecnología de reconocimiento de voz 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. De acuerdo a la literatura, con un sistema ASR

que entregue una exactitud cercana a 90 % estos sistemas tienden a ser muy robustos [48]. Por otra parte, los sistemas de respuesta abierta son mucho más complejos y exigentes en cuanto a requerimientos técnicos y pedagógicos.

La tecnología ASR también ha sido aplicada al problema de la evaluación 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 [49]. Hamid y Rashwan [50] y Abdou *et al.* [51] proponen una técnica donde el modelo de lenguaje el reconocedor se genera tomando en consideración errores como eliminación y sustitución de palabras, o eventuales errores de pronunciación. La medida de confiabilidad de la pronunciación se basa en la duración de los fonemas. Moustroufas y Digalakis [52] utilizaron dos reconocedores en paralelo para combinar modelos de inglés nativo con no nativo. Molina *et al.* [53] proponen una técnica de evaluación de pronunciación basado en ASR usando modelos competitivos. El método propuesto emplea además fusión de múltiples clasificadores (MCS, *multi-classifier systems*) para combinar los *scores* que provienen de distintas fuentes de información. Muchas de las técnicas encontradas en la literatura especializada muestran correlaciones que varían entre 0,6 y

0,8 [54, 55, 56, 57, 58, 59, 60, 61].

### 2.6.1. Evaluación automática de prosodia

Existen muchas técnicas que combinan la pronunciación de sonidos individuales (segmentos) con prosodia (suprasegmentos). Neumeyer *et al.* explica que los estudiantes de idiomas tienden a hablar más lento la lengua que están aprendiendo [62]. Para solucionar este problema incorpora el *speaking rate* (medida de la cantidad de palabras por unidad de tiempo) en el puntaje de pronunciación. Dong *et al.* aplican estrategias que combinan las medidas calculadas a nivel de segmento con otras extraídas F0, energía y duración [63]. Además, muchos sistemas comerciales generan un único *score* a partir de características segmentales y prosódicas. No obstante, con el enfoque anterior el usuario no puede distinguir la prosodia de la calidad de pronunciación, por lo que se puede generar confusión a los estudiantes que recién comienzan a estudiar un idioma y a los de niveles más básicos. Por tanto, es lógico que los aspectos segmentales y suprasegmentales de la pronunciación en enseñanza de idiomas sean analizados por separado.

El problema de la evaluación de la evaluación de prosodia ha sido ha sido abordada en la literatura desde varios puntos de vista. Tepperman y Narayanan proponen métodos texto independientes para medir el grado de *nativeness* (es decir, qué tan nativo es un usuario determinado) a través del análisis de la frecuencia fundamental [64]. Eskenazi *et al.*, presentan una estrategia de entrenamiento de fluidez basado en la

medición de características prosódicas [65]. Este sistema invita al usuario a repetir una frase dada, y de esta forma puede entrenar distintos aspectos como duración, pitch, etc. La información de la duración está dada por el algoritmo de Viterbi forzado. Cabe destacar que este algoritmo puede estimar automáticamente los inicios y fines de los sonidos dada la transcripción de la elocución. Peabody *et al.* presenta un método automático de corrección de tono para el chino mandarín de hablantes no nativos [66]. El sistema compara un modelo con los contornos de F0 generados por los usuarios.

No obstante lo anterior, el problema de evaluación de prosodia desde el punto de vista de entonación en enseñanza de segundo idioma no ha sido exhaustivamente abordado en la literatura. La mayoría de los trabajos se centran en el problema de la pronunciación a nivel de segmentos, es decir, en la evaluación de calidad de la pronunciación [62, 67, 49]. Sin embargo, algunos autores han utilizado la entonación como una característica adicional al problema de evaluación de pronunciación [63]. En [68], se presenta un módulo de prosodia en enseñanza de segundo idioma, el cual incluye actividades de entonación y acento. El sistema compara la señal de referencia con la señal del alumno utilizando una heurística. La desventaja de este sistema necesita humanos para ingresar información ortográfica a la referencia y además no entrega ningún puntaje a partir de la comparación. En [69, 70] se propone un método de evaluación de calidad de entonación basado en una filosofía bottom-up, donde la entonación es clasificada sílaba a sílaba. El sistema utiliza alineamiento de Viterbi forzado, y por tanto es texto dependiente. Finalmente, en [71] se propone un método para el diagnóstico de

desórdenes del habla y el lenguaje. El sistema alcanza altos valores de correlación con las evaluaciones subjetivas, no obstante necesita la transcripción y las etiquetas de los segmentos fonéticos.

La mayoría de las técnicas propuestas en la literatura se basan en la filosofía *bottom-up* (es decir, de abajo-arriba). Esto significa que el análisis se efectúa analizando pequeños trozos para luego combinarlos. Por el contrario, una filosofía *top-down* (arriba-abajo), el análisis se formula en forma general, y luego se efectúa una exploración más segmentada. En lingüística, el análisis de sonidos y palabras ha pasado a un análisis que involucra unidades más grandes, como textos completos, discursos e interacciones, originando nuevas disciplinas como el análisis de discurso, pragmática, análisis de conversaciones y de discurso [72, 73]. En enseñanza de idiomas, la tendencia actual es centrarse en la efectividad de la comunicación y no en la pronunciación exacta de los segmentos o sonidos individuales, lo cual implica además incorporar las características suprasegmentales. En este sentido se está adoptando la filosofía *top-down*, haciendo énfasis en la comunicación y en el significado global más que en sonidos aislados. Sin embargo, la superioridad de *top-down* versus *bottom-upes* aún un tema de debate más que una realidad aceptada.

### **2.6.2. Medidas de desempeño en CALL**

En evaluación de pronunciación y entonación la medida utilizada para medir el desempeño es la correlación entre los *scores* subjetivos (que resultan a partir de eva-

luaciones de expertos) y objetivos (dados por el sistema). Dado un conjunto de señales, la correlación se calcula de acuerdo con:

$$\rho = \frac{Cov(score_{subetivo} - score_{objetivo})}{Var(score_{subetivo}) \cdot Var(score_{objetivo})} \quad (2.24)$$

A mayor correlación, mejor es el desempeño del sistema de evaluación de entonación. Además, esta medida también es aplicable para un sistema de evaluación de velocidad de lectura.

En el caso de de evaluación de acento léxico o *stress* se utilizan medidas propias de un sistema de clasificación. Evaluar el acento léxico es un problema de dos clases: el acento del usuario es correcto; o bien es incorrecto. De esta forma, se tienen cuatro casos posibles: clasificar el acento como correcto dado que es correcto; como incorrecto dado que es incorrecto; como correcto dado que es incorrecto; y, como incorrecto dado que es correcto. Los dos primeros casos corresponden a respuestas correctas del sistema, mientras que los últimos dos casos son erróneos y se denominan “falso positivo” (FP) y “falso negativo” (FN). Dado un umbral de decisión o una determinada configuración, el valor que iguala ambos errores se denomina Equal Error Rate (EER), que ha sido ampliamente usado en sistemas de reconocimiento de patrones. Otra medida de desempeño utilizada es la curva ROC (Receiver Operating Characteristic), la que consiste en graficar la tasa de FN versus (1-FP) para un amplio rango de umbrales de decisión. El área bajo esta curva es utilizada como indicador de la capacidad discriminativa del sistema a evaluar. A menor área bajo la curva ROC, mejor es el desempeño.

## 2.7. Reconocimiento y detección de emociones en señales de voz

### 2.7.1. Antecedentes

De acuerdo a Cowie *et al.*, en la interacción humana existen dos canales de comunicación: el explícito y el implícito [4]. El primero hace referencia al mensaje expreso entregado por un individuo. Por su parte, el canal implícito indica cómo interpretar el canal explícito. En otras palabras, el canal implícito define la capacidad que tienen los seres humanos de manejar, reconocer y comprender las emociones. Por ejemplo, la sentencia “Se nota que has aprendido bastante en el colegio” puede ser interpretada de forma literal (mensaje explícito) pero también, en determinada circunstancia, podría significar exactamente lo contrario de acuerdo a la *forma* en que es pronunciada (canal implícito).

El estudio del canal implícito ha sido abordado por diversas disciplinas, principalmente por la lingüística y la psicología. No obstante, dada su alta complejidad, aún es materia de investigación. En virtud de sus variadas aplicaciones, el estudio de las emociones ha despertado el interés de áreas como la ingeniería. Como ha sido explicado anteriormente, la detección, reconocimiento y síntesis de emociones tiene aplicaciones en interfaces hombre-máquina sensibles a emociones que mejoren la experiencia del usuario. Hay sistemas que requieren indicadores emocionales objetivos para hacer juicios sobre un individuo, por ejemplo, en la detección de mentiras o en la diagnóstico

de enfermedades mentales. Otra aplicación interesante es la generación de alertas a partir de estados emocionales de un usuario. Por ejemplo, detectar a una enfermera alterada por una situación complicada o redirigir a un cliente molesto con un operador entrenado en un *call center*. La industria del entretenimiento también se ha interesado en desarrollar juegos que logren responder al estado emocional del usuario.

La emoción es percibida por los humanos mediante los sentidos, y es transmitida por expresiones faciales [74], gestos [75] y la voz [76, 77, 78]. En la señal de voz, la prosodia es especialmente relevante dado que las emociones se expresan a través de cambios en la entonación, en la cantidad de sílabas por segundo o *rate*, la intensidad y la duración. En la literatura se pueden encontrar diversos trabajos que han estudiado la relación entre parámetros acústicos de la voz con el estado emocional. Por ejemplo, Fonagy mostró que en emociones como enojo, felicidad el valor promedio de F0 aumenta, mientras que en el caso de sadness disminuye [79]. Van Bezooijen realizó un estudio similar para la energía, en el que se reportan aumentos y disminuciones para los estados afectivos felicidad y tristeza, respectivamente [80]. La emoción también se presenta a nivel de segmentos. Goudbeek *et al.* analizaron la posición de los formantes en función del estado emocional [81]. Sus resultados mostraron que el primer y segundo formante toman valores más altos que el promedio para emociones con alta excitación (*arousal*) y valencia positiva, respectivamente.

## 2.7.2. Representación de las emociones

El objetivo de un reconocedor de emociones es asignar una etiqueta que identifique el estado emocional de un individuo. Sin embargo, definir las etiquetas no es una tarea trivial. Por ejemplo, Cowie *et al.* muestra alrededor de 130 emociones diferentes [4]. Lógicamente, es muy difícil que un sistema automático logre ese nivel de discriminación. Por esta razón, en la literatura se utilizan otras representaciones del estado afectivo de un individuo, muchas de ellas inspiradas en la psicología. Para este trabajo es de especial interés la representación activación-evaluación (*activation-evaluation*) que se explica a continuación.

El espacio activación-evaluación permite representar de manera simple un amplio rango de emociones y consiste en un plano cartesiano donde el eje  $y$  indica la activación y el eje  $x$  la valencia. El nivel de activación o *activation* determina cuán dinámica es la emoción, esto es, si es activa o pasiva. Por ejemplo, los estados emocionales “felicidad” y “tristeza” se caracterizan por ser activos y pasivos, respectivamente. Por otra parte, la dimensión evaluación (*evaluation*), también denominada valencia o *valence*, indica si la emoción es negativa o positiva. Por ejemplo, tanto “felicidad” como “enojo” son emociones activas. Su diferencia radica en que la primera es positiva mientras que la segunda es negativa. La Figura 2.6 muestra gráficamente la representación activación-evaluación.

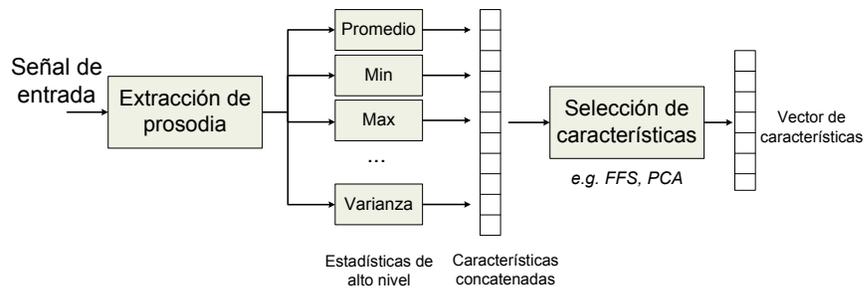


**Figura 2.6:** Representación “activación-evaluación”. El eje horizontal diferencia las emociones positivas y negativas, mientras que el eje vertical discrimina entre estados emocionales activos y pasivos.

### 2.7.3. Técnicas estándar en reconocimiento de emociones en señales de voz

En virtud de la evidencia en la literatura sobre la relación entre el estado afectivo y los parámetros acústicos de la señal de voz, el enfoque más usado en reconocimiento de emociones consiste en extraer estadísticas de alto nivel o funcionales (media, varianza, kurtosis, mínimo, máximo, y rango) a partir de los parámetros prosódicos, fundamentalmente F0 y energía (así como también su primera y segunda derivada) [82, 83]. Este proceso usualmente se aplica a nivel de señal (i.e. frase completa u oración). Las estadísticas se concatenan para formar un vector de parámetros que identifica el estado emocional de la elocución. Luego, se emplean técnicas como *forward feature selection* (FFS), algoritmos genéticos, LDA (*linear discriminant analysis*), análisis de

componentes principales (PCA, *principal component analysis*) para reducir el número de funcionales y escoger aquellos que sean emocionalmente más relevantes [77, 84, 85]. Finalmente, dada una base de datos emocional, se entrena un clasificador para discriminar entre dos o más clases emocionales (por ejemplo, tristeza versus alegría). La Figura 2.7 muestra un diagrama en bloques del enfoque descrito anteriormente.



**Figura 2.7:** Diagrama en bloques del enfoque estándar en detección de emociones.

Si bien las técnicas basadas en estadísticas de parámetros acústicos a nivel de señal han entregado buenos resultados, no han estado exentas de críticas. En primer lugar, al calcular estadísticas a nivel global se descarta implícitamente información valiosa a nivel de sub-oración [86, 87]. Esta discusión fue propuesta por Busso *et al.* y en su trabajo se realizaron experimentos a nivel de *voiced-level* (i.e. se utilizan los segmentos sonoros como unidad de análisis). Por otra parte, no se toma en cuenta la forma de la frecuencia fundamental o la energía en función del tiempo, pese a que algunos autores han afirmado que su tendencia y forma contienen información emocional relevante [88, 89]. En efecto, si una señal es muy larga y la información afectiva está concentrada

en un segmento determinado, la parte no-emocional eventualmente podría enmascarar dicho segmento.

#### 2.7.4. Avances recientes

Busso *et al.* introdujeron el concepto de usar un modelo neutral para clasificar emociones [77]. La idea es contrastar la elocución de test con modelos neutrales con el objeto de clasificar binariamente a la señal acústica de entrada como emocional o neutra. Para efectuar este procedimiento, basta entrenar el modelo usando solamente muestras de voz emocionalmente neutras. Este enfoque tiene una serie de ventajas cuando se compara con el método convencional. En primer lugar, como se menciona más arriba, es difícil establecer con exactitud las clases emocionales. Segundo, la disponibilidad de bases de datos emocionales es reducida, y por tanto no es posible entrenar modelos robustos. A esto se agrega el riesgo de sobreentrenamiento al utilizar un corpus emocional específico, hecho que hace impracticable el uso de estos sistemas en aplicaciones reales. Dado que existe un número considerable de bases de datos neutrales disponibles, entonces es posible construir modelos robustos que pueden ser usados para discriminar entre neutro y emocional. Además, este módulo puede funcionar como primera etapa para luego discriminar entre clases específicas definidas de acuerdo a la aplicación. Cabe destacar que existen varias aplicaciones donde interesa determinar con exactitud si un usuario está hablando de forma neutra o no (e.g. *call center*).

Algunos trabajos han representado la forma de los parámetros prosódicos con el

fin de capturar información que se pierde al estimar estadísticas globales. Rotaru y Litman utilizaron características de F0 a nivel de palabras y de oración incorporando características como el coeficiente de regresión para aproximar la forma del contorno de F0 [90]. Este coeficiente permite representar la dirección del contorno de pitch (i.e. si es ascendente o descendente). Los autores mencionados también propusieron usar coeficientes de segundo orden para modelar concavidad/convexidad de las curvas. Busso *et al.* también incorporan en su trabajo características similares para modelar la forma del contorno de F0 como la pendiente, la curvatura y la inflexión basados en la investigación de Grabe *et al.*, quien propone descriptores para modelar la forma de la frecuencia fundamental usando polinomios [91].

Finalmente, cabe mencionar que también hay investigadores que han desarrollado sistemas automáticos de reconocimiento de emociones basado en parámetros espectrales como LPC (*linear prediction coefficients*), MFCC (*Mel-frequency Cepstral coefficients*) MFB (*Mel filter bank*) o [92, 93, 94]. En forma alternativa a las estadísticas de alto nivel también se han utilizado técnicas como HMM para identificar la emoción en una secuencia de cuadros [95].

### **2.7.5. Medidas de desempeño en reconocimiento y detección de emociones**

Al igual que en el caso de enseñanza de idiomas, para medir el desempeño de sistemas de detección de emociones es necesario contar con etiquetas o evaluaciones de

referencia generadas por expertos que sirvan de *ground truth*. Una opción es etiquetar cada señal de acuerdo a un estado emocional determinado (e.g. neutral, alegría o ira). En este caso se pueden utilizar como medidas de desempeño el *accuracy* (exactitud), *precision*, *recall* y *F-measure* definidas de acuerdo a:

$$\text{Accuracy} = \frac{VP + FN}{VP + FP + VN + FN} \quad (2.25)$$

$$\text{Precision} = \frac{VP}{VP + FP} \quad (2.26)$$

$$\text{Recall} = \frac{VP}{VP + FN} \quad (2.27)$$

$$\text{F-measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.28)$$

donde *FP*, *FN*, *VP* y *VN* denotan las tasas de falsos positivos, falsos negativos, verdaderos positivos y verdaderos negativos, respectivamente.

El esquema descrito anteriormente es utilizado en bases de datos emocionales actualizadas. Sin embargo, no es aplicable a elocuciones afectivas reales dado que la emoción no se distribuye de forma constante en el tiempo [92, 96]. En este caso, es necesario contar con una medición de la emoción en función del tiempo. Cowie *et al.* presentaron una herramienta que permite etiquetar la emoción en tiempo real llamada Feeltrace [97]. Este sistema permite a evaluadores etiquetar material audiovisual en el espacio *activation-evaluation* en tiempo real mediante una interfaz gráfica. Como resultado, se obtienen las curvas  $a(t)$  y  $v(t)$  para la activación y valencia, respectivamente, en

función del tiempo (ver sección 2.7.2). Por lo tanto, es posible calcular el desempeño en reconocimiento de emociones de acuerdo a la correlación entre las mediciones subjetivas dadas por el evaluador y las objetivas entregadas por un sistema automático determinado.

## Capítulo 3

# Evaluación de entonación y acento en enseñanza de idiomas

En este capítulo se discute la relevancia de la información que provee la entonación en el marco de la enseñanza de segundo idioma. Como consecuencia, se propone un sistema para la evaluación automática de la entonación para enseñanza de segundo idioma basado en un esquema *top-down*. Además, se presenta un sistema para la evaluación del acento léxico o *stress* que combina la información de la frecuencia fundamental y la energía. La elocución pronunciada por el estudiante es directamente comparada con una referencia. La similitud de la entonación y de la energía son estimadas *frame a frame* usando el algoritmo DTW. Además, se evalúa la robustez del alineamiento entregado por el algoritmo DTW al micrófono, speaker y calidad de pronunciación. El sistema para la evaluación de la entonación entrega un *score* de correlación entre mediciones

objetivas y subjetivas igual a 0.88. Por su parte, el sistema para evaluar acento léxico da un EER (*Equal Error Rate* igual a 21.5 %, que a su vez es similar al error observado en esquemas de evaluación fonética. Estos resultados sugieren que los sistemas propuestos pueden ser usados en aplicaciones reales. Finalmente, los sistemas presentados aquí son independientes del texto y del idioma debido a que no se requiere la transcripción ni información del idioma de la elocución.

### **3.1. Introducción**

La enseñanza de idiomas asistida por computador (*CALL, Coputer aided language learning*) ha reemplazado los paradigmas tradicionales (por ejemplo audios de laboratorio) con interfaces hombre-máquina que proveen interacciones más naturales. Los antiguos sistemas basados en ilustraciones estáticas son reemplazados por diálogos reales entre el usuario y el sistema, donde es posible evaluar la pronunciación o la calidad de la fluidez y dar respuestas por voz. En este nuevo paradigma, las tecnologías de voz han jugado un rol muy importante. Como resultado, los sistemas CALL ofrecen muchas ventajas a los estudiantes y el proceso de aprendizaje se desarrolla en un contexto altamente motivante caracterizado por la interactividad [98]. Normalmente, los estudiantes se sienten incómodos e inhibidos al hablar en una sala de clases [99]. Por lo tanto, los sistemas CALL pueden proveer un entorno más conveniente para practicar un segundo idioma.

Las características suprasegmentales de la voz como el pitch, el volumen y la velo-

cidad [14] son de suma importancia cuando se enseña un segundo idioma. Por ejemplo, la mayoría de los estudiantes de inglés puede alcanzar un nivel aceptable de habilidades de escritura y lectura, pero su pronunciación rara vez alcanza el mismo estándar. Problemas habituales son la falta de fluidez y naturalidad, entre otros. Cabe destacar que para algunos autores naturalidad de estilo implica fluidez. Por ejemplo, de acuerdo a [100], “el grado de aislamiento de contexto, o incluso el tipo de texto en sí mismo, puede evocar distintos grados de naturalidad en el estilo, y por tanto en la fluidez”. Además, algunas veces los profesores muestran muy bajo nivel de habilidades orales [49, 101], lo que a su vez es una barrera adicional para estudiantes principiantes. A pesar de que las reglas fonéticas (entendidas como reglas para la correcta pronunciación de segmentos [102, 103, 104]) concentran la mayoría de la atención en el proceso de aprendizaje de habilidades de comunicación oral, en el caso de los estudiantes avanzados, la prosodia es uno de los aspectos más importantes [68] para alcanzar una pronunciación natural y fluida comparada con hablantes nativos. En este contexto, el análisis de la señal de voz es muy importante para ayudar a los estudiantes a practicar y mejorar sus habilidades orales, sin la necesidad de la asistencia de un profesor [105]. Además, proveer retroalimentación adecuada es un tema muy relevante en CALL [10] porque éste puede motivar a los estudiantes a mejorar y practicar su pronunciación. En la literatura existe evidencia que el feedback audiovisual puede mejorar la eficiencia del entrenamiento de la entonación [9, 106].

En el contexto de la prosodia y la enseñanza de segundo idioma, la entonación es más

importante que la energía y la duración. La entonación está fuertemente relacionada con la naturalidad, la emoción e incluso el significado, como se explica más adelante en la sección 2 de este capítulo. Además, los acentos en las palabras son el resultado de movimientos de F0 los cuales juegan un rol en el mecanismo del *stress* silábico [70]. El problema de la entonación ha sido abordado desde varios puntos de vista: medición del grado de *nativeness*; evaluación y entrenamiento de fluidez; clasificación; y, evaluación asistida por computador de la calidad de la pronunciación. En [64, 107] se utilizan métodos texto independiente para evaluar el grado de *nativeness* analizando la curva de F0. En [65] se presenta una estrategia para entrenar la fluidez a través de la medición de características prosódicas. En este método se solicita al usuario repetir una frase u oración dada. Luego, el sistema corrige la duración separada de otras características. Finalmente, el usuario procede con el pitch. La información de la duración es entregada por el algoritmo forzado de Viterbi [108]. Vale la pena mencionar que el algoritmo forzado de Viterbi puede estimar automáticamente los límites de los fonemas dada una señal y su transcripción. En [66] se presenta un sistema para corregir automáticamente el tono en mandarín no nativo. La técnica compara modelos tonales independientes del locutor con los contornos de pitch generados por los usuarios. Notar que en [64, 107, 65, 66] se utiliza la filosofía *bottom-up* para evaluar las características prosódicas usando modelos independientes del texto o del locutor. Además, observar que el problema de la evaluación de entonación del punto de vista de CALL no es necesariamente una medición del nivel de *nativeness*, evaluación de fluidez o clasificación del contorno de

pitch con clases predefinidas.

Sorprendentemente, el problema de la evaluación de la calidad de la pronunciación del punto de vista de la entonación en CALL no ha sido abordado exhaustivamente en la literatura. La mayoría de los trabajos en evaluación de pronunciación se han centrado en la calidad fonética [62, 67, 49]. Sin embargo, algunos autores han usado la entonación como variable adicional para medir la calidad de la pronunciación en combinación con otras características [63]. En [68], se presenta un módulo prosódico para enseñanza de idiomas que incluye actividades de entonación y acento léxico. El sistema contrasta la señal del estudiante con una referencia usando heurísticas. El método requiere asistencia humana para insertar información ortográfica y además no entrega ningún tipo de puntaje o *score*. En [69, 70] se propone un sistema de evaluación de la entonación basado en un esquema bottom-up donde se clasifica la curva de F0 de cada sílaba. El sistema usa alineamiento forzado de Viterbi y por tanto es texto dependiente. En [71] se propone un método de medición prosódica para el diagnóstico de desórdenes del habla. Los resultados muestran una alta correlación entre las evaluaciones automáticas y aquellas realizadas por expertos. Sin embargo, el sistema requiere la transcripción del texto o bien la segmentación fonética.

Las reglas fonéticas pueden ser fácilmente clasificadas como “correctas” e “incorrectas” de acuerdo a una ubicación geográfica. Por el contrario, normalmente hay más de un patrón de entonación que puede ser considerado “aceptable” dado un texto [109]. Esto se debe a que la entonación provee información de las emociones, intenciones y

actitudes. Como resultado, en vez de clasificar la curva de entonación como correcta o incorrecta, es más conveniente motivar al estudiante a seguir un patrón de referencia dado.

En este trabajo se presenta un sistema automático para detectar entonación basado en un esquema *top-down*. La técnica propuesta intenta separar la evaluación de entonación de la calidad de pronunciación del estudiante. Dada una señal de referencia, el estudiante puede escuchar y repetir una elocución dada imitando el patrón de entonación de referencia. Después, las señales de referencia y test son alineadas frame a frame usando *dynamic time warping* (DTW). Se estima el pitch en ambas señales y luego se aplica post procesamiento para eliminar errores de *halving* o *doubling* en el cálculo de la frecuencia fundamental. Los contornos de pitch de referencia y test resultantes se representan en la escala de semitonos y se normalizan de acuerdo a su media. Luego, la medida de similitud entre la señal de referencia y de test es evaluada *frame a frame* usando el alineamiento DTW mencionado anteriormente. En vez de calcular la diferencia entre las curvas de F0 de referencia y de test, este trabajo propone estimar la correlación entre las curvas. Finalmente, el stress a nivel de sílabas se mide usando la información del pitch en conjunto con la energía a nivel de frame. El sistema propuesto es texto independiente (es decir, no es necesario contar con la transcripción de la señal de referencia), minimiza el efecto de la calidad de pronunciación a nivel de segmentos del usuario y entrega una correlación de *scores* subjetivos (entregados por evaluadores expertos) y objetivos (dados por el sistema propuesto) igual a 0.88 para evaluación de

entonación. La evaluación del *stress* a nivel de palabra, que resulta de una combinación del contorno de pitch y la energía, entrega un *equal error rate* (EER) igual a 21.5 %, el que a su vez es comparable al error de los sistemas de evaluación de pronunciación a nivel de segmento. Pese a que el sistema propuesto en este trabajo es probado en lengua inglesa, puede ser considerado independiente del idioma. Las contribuciones de este trabajo son: (a) una discusión del rol de la entonación en enseñanza de idiomas; (b), un sistema texto independiente para evaluar la entonación; (c), el uso de la correlación para comparar las curvas de entonación; (d), un sistema texto independiente para medir el acento léxico enseñanza de segundo idioma; y, (e) una evaluación de la robustez del alineamiento DTW respecto al locutor, pronunciación de segmentos y *mismatch* de micrófono.

## **3.2. La importancia de la entonación en enseñanza de segundo idioma**

### **3.2.1. Definiciones**

Una descripción fonética adecuada estaría incompleta si no se tiene en cuenta algunas características de gran importancia que acompañan a los segmentos. Estas características se conocen como elementos suprasegmentales. Los más importantes son el pitch, la intensidad y la duración [13]. De acuerdo a este autor, el pitch es la percepción de la frecuencia fundamental, la manifestación acústica de la entonación; lo que

es intensidad en el extremo receptor debe estar relacionado con la intensidad en la fase de producción, que a su vez se relaciona con el tamaño o la amplitud de la vibración; y, la duración está relacionada con la el largo de un segmento, aunque algunas veces “las variaciones del largo en términos acústicos pueden no corresponder a nuestros juicios lingüísticos de duración.

### **3.2.2. Entonación**

De acuerdo a Botinis *et al.* [9], “la entonación se define como la combinación de características tonales en unidades estructurales más grandes asociadas a la frecuencia fundamental F0 y sus variaciones distintivas en el proceso del habla. F0 se define por el número de ciclos cuasiperiódicos por segundo de la señal de voz y se mide en Hz”. De hecho, F0 corresponde al número de veces por segundo que las cuerdas vocales completan un ciclo de vibración. En consecuencia, la producción de la entonación está regulada por las fuerzas musculares de la laringe que controlan la tensión de las cuerdas vocales, además de las fuerzas aerodinámicas del sistema respiratorio. El pitch percibido, que corresponde aproximadamente a F0, define la percepción de la entonación.

La entonación tiene muchas funciones pragmáticas de importancia [10, 11]. En este punto es necesario decir que siempre va acompañada de otros rasgos suprasegmentales como la intensidad y la duración. Entre sus muchas funciones, se puede decir que la entonación es particularmente relevante para expresar actitud, prominencia, relaciones gramaticales, para estructurar el discurso y dar naturalidad al habla [12, 13, 14].

Las emociones y actitudes que utilizan las personas cuando hablan se reflejan en la entonación. La misma frase puede mostrar diferentes actitudes dependiendo de la entonación con la que se pronunció. Ésta es la función actitudinal o expresiva de la entonación. Además, juega un rol importante en la asignación de protagonismo a las sílabas que debe ser reconocidas como acentuadas. Esta función se suele llamar acentual. La entonación también tiene una función gramatical, ya que proporciona información que hace más fácil para el interlocutor a reconocer la estructura gramatical y sintáctica de lo que se dice, como por ejemplo determinar la ubicación de la frase, cláusula o límite de oración, o la distinción entre construcciones interrogativas y afirmativas. Esta función comúnmente se denomina gramatical. Teniendo en cuenta habla desde una perspectiva más amplia, la entonación puede sugerir al oyente lo que tiene que ser considerado como “nueva” información y lo que se considera como algo “dado”. También puede sugerir qué es lo que el locutor indica como cambio o vínculo con algún material presente en otra unidad tonal y, en una conversación, puede proporcionar una sugerencia en relación con el tipo de respuesta que se espera. Ésta es la función del discurso de la entonación. La última función es difícil de describir, pero es fácilmente reconocible por todo hablante nativo competente. Tiene que ver con el resultado del uso de la entonación adecuada, lo que proporciona a la naturalidad del habla. Esto puede estar relacionado con la función *indexical* que describe Wells [14] “... la entonación puede actuar como un marcador de identidad personal o social. Lo que hace que las madres suenen como madres, los amantes como amantes, y los abogados como abogados, ...”

Un hablante nativo puede reconocer sin gran esfuerzo si un enunciado ha sido producido por un hablante nativo o no. Hay muchas características que contribuyen a este objetivo, algunos de los cuales son más fáciles de distinguir que los demás: la elección de palabras, la estructura sintáctica, las características segmentales, y, sin duda, la entonación. Sin embargo, podría suceder a un hablante competente de una lengua extranjera que, si su entonación no es exactamente la que un hablante nativo utilizaría en las mismas circunstancias, su discurso se oiga artificial y llame la atención la forma en que lo dijo y no su contenido.

### 3.2.3. Acento léxico

Algunos autores evitan el uso de la palabra “*stress*”, ya que, como lo indica Crutenden [13], este término se emplea en la fonética y la lingüística de diversas formas: a veces se utiliza como un equivalente a la intensidad, a veces como “lo que hace prominente y que sea distinto al pitch” (es decir, por la intensidad o la duración), y, en ocasiones, se refiere a las sílabas de las unidades léxicas que tienen el acento. Este trabajo usa la definición presentada por Wells [14]: “el *stress* o acento léxico es una combinación de intensidad, pitch y duración”.

En una palabra como *mother*, el acento recae en la primera sílaba. En *university*, la sílaba “*ver*” recibe el acento primario, mientras que la sílaba “*u*” recibe un acento secundario. Las sílabas “*ni*” “*si*” y “*ti*” se consideran no acentuadas. La presencia de sílabas que reciben un acento principal o secundario es importante en inglés ya que

a nivel de segmento tienden a ser pronunciadas en su totalidad. El debilitamiento de las vocales y la reducción vocálica por lo general ocurren en las sílabas no acentuadas. La importancia del acento secundario radica en este hecho, es decir, que la reducción vocálica es el resultado de algunas sílabas sin acento. Cabe destacar que, en muchos otros idiomas distintos al inglés como el italiano o español, el acento secundario no afecta la pronunciación de segmentos. Sin embargo, es una práctica común en enseñanza de idiomas la de centrar la atención en el acento primario [110], el cual afecta al significado de una determinada palabra. Posicionar mal el acento secundario puede afectar a la pronunciación de los segmentos, pero no necesariamente al significado. Por otra parte, por razones de viabilidad, las palabras objetivo en los experimentos de este trabajo fueron elegidos con el fin de evitar el *stress* secundario. A pesar de que el acento secundario es un tema relevante en la adquisición del lenguaje en los niveles avanzados, esta investigación se centra en el acento primario.

### **3.2.4. La importancia de la entonación**

#### **3.2.4.1. La importancia de la entonación en general**

Como se ha indicado anteriormente en este capítulo, la prosodia es muy importante. La entonación es central en el proceso de comunicación [111]. Los hablantes de cada idioma reconocen el rol que juega la prosodia cuando se hacen comentarios como: “Estuvo de acuerdo, pero lo dijo de tal manera ...” En muchas ocasiones esta “manera” en que se dice algo es más importante que el mensaje literal, la organización sintáctica

o las palabras utilizadas para la estructura [112]. Con más frecuencia de lo que se puede imaginar, las características prosódicas puede sugerir precisamente el sentido contrario que las palabras reales usadas por el locutor. La entonación es tan importante que incluso puede ser utilizada sin decir una palabra. Un sonido único, por ejemplo, /m/, se puede decir con diferentes entonaciones indicando acuerdo, duda, desacuerdo, placer, crítica, entre otras actitudes [113, 114]. No es de extrañar que éste es uno de los primeros aspectos del lenguaje a los que un niño presta atención, reacciona y produce él mismo. De acuerdo a Peters [115] citado por Cruttenden [13]: “muchos bebés son excelentes imitadores de la entonación y pueden producir un sonido con patrones de entonación característicos de la lengua inglesa en sílabas sin sentido durante la última etapa de su balbuceo pre-lingüístico”. Además, existe una estrecha relación entre la prosodia y la sintaxis. Como explica Wells [14] “La entonación ayuda a identificar las estructuras gramaticales en el habla, tal como los signos de puntuación lo hacen por escrito”.

#### **3.2.4.2. La importancia de la entonación en la enseñanza de segundo idioma**

A pesar de que muchos lingüistas hablan de la entonación en lenguas diferentes como si fueran entidades discretas, dentro de un mismo idioma existen varios sistemas de entonación diferentes [116, 112]. Un hablante nativo puede muy fácilmente, y sin ningún entrenamiento previo, detectar que otro hablante nativo del mismo idioma está utilizando un dialecto diferente. Para ello, dicho hablante detecta patrones de entonación con los cuales no está familiarizado. De acuerdo con [117], “Existen con-

siderables diferencias entre los patrones de entonación que se encuentran en todo el mundo de habla española. Incluso dentro de un área geográfica relativamente pequeña puede haber grandes diferencias de entonación”. En este marco, comparar la entonación del inglés y del español es una tarea imposible. No obstante, sí se podría comparar la entonación de dos dialectos determinados de uno de estos idiomas. A pesar de que hay diferencias de entonación dentro de una lengua, hay algunas características que son comunes a muchos idiomas. De acuerdo a Wells [14], “Al igual que otras características prosódicas, la entonación en parte es universal, pero también en parte es específica del idioma”. Así, en varios idiomas una entonación descendente se asocia con una declaración o una orden, y una melodía ascendente, con una declaración incompleta, una pregunta o una petición amable. Sin embargo, hay diferencias que podrían dar lugar a malentendidos, sobre todo en lo que respecta a las intenciones o actitud del locutor, que puede sonar, por ejemplo, grosero o insistente en vez de cortés o amable. Existe evidencia empírica que muestra la existencia de diferencias significativas en la elección de la entonación y del acento de pitch en hablantes de inglés no nativos que pueden causar malentendidos de comunicación [118]. A pesar de que un hablante no nativo puede utilizar la entonación correcta, el problema podría estar en el hecho de que el núcleo está fuera de lugar, donde el núcleo corresponde a la sílaba identificada por el acento de pitch [13]. Se sabe que en lenguas como el francés, italiano y español el núcleo está en la última palabra en la frase de entonación, lo que no necesariamente ocurre en inglés. En consecuencia, errores como acentuar la palabra “*it*” en lugar de “*thought*” en

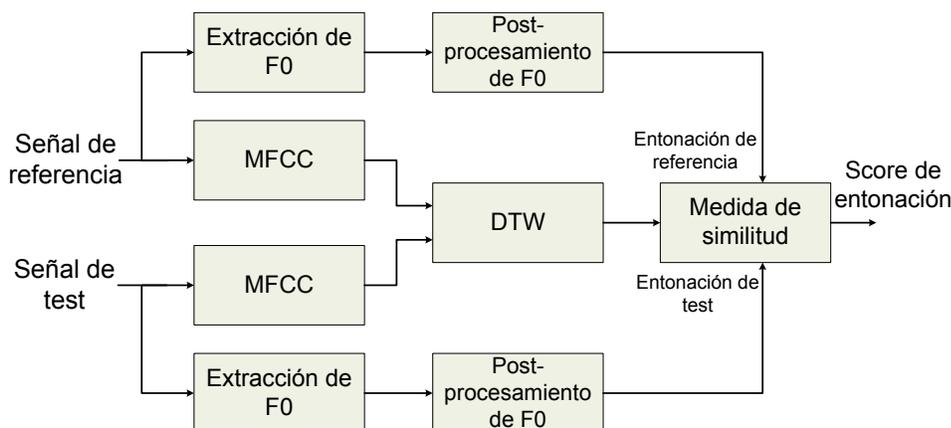
*“I haven’t thought about it”* se escucha con frecuencia [13, 14]. Los hablantes nativos del inglés pueden distinguir fácilmente las desviaciones gramaticales, léxicas y de pronunciación producidas por hablantes no nativos, y por lo tanto tener en cuenta sus errores. Sin embargo, son incapaces de hacerlo con la entonación. De acuerdo a Wells [14], “Los hablantes nativos de inglés saben que los estudiantes de este idioma tienen dificultades con las vocales y consonantes. Al interactuar con alguien que no es un hablante nativo de inglés, un hablante nativo puede tolerar los errores a nivel de segmentos, pero no puede aceptar errores de entonación. Probablemente, esto se debe a que no es capaz de darse cuenta que la entonación puede ser errónea”. La lingüística tradicional ha ampliado su estudio de sonidos, palabras y frases a unidades más grandes tales como textos completos, discursos e interacciones, dando lugar a disciplinas como el análisis del discurso, la lingüística del texto, la pragmática y el análisis conversacional [72, 73]. En la actualidad, la lingüística aplicada hace hincapié en la importancia de la entonación, que junto con el *stress* y el ritmo, no sólo complementan el significado, sino que lo crean [10, 13, 119, 120]. Por esta razón, en la actualidad la enseñanza de idiomas se centra en la eficacia comunicativa y, en consecuencia, se ha dado mayor importancia a las “características suprasegmentales en vez de los sonidos individuales” [119]. En otras palabras, hay una tendencia a adoptar un esquema *top-down*, es decir, a enfocarse más en la comunicación y el significado global en vez de atenerse al enfoque tradicional *bottom-up* centrado en los sonidos aislados [121, 122, 123, 124]. Sin embargo, vale la pena mencionar que la superioridad del enfoque *top-down* sobre el esquema *bottom-up*,

o viceversa, es aún materia de debate.

### 3.3. El sistema propuesto

El sistema propuesto intenta decidir, mediante un enfoque *top-down*, si dos elocuciones (es decir, de referencia y de test), provenientes de distintos locutores, fueron pronunciadas con el mismo patrón de entonación. La Fig. 3.1 muestra un diagrama de bloques del sistema propuesto. En primer lugar, se estima la frecuencia fundamental F0 y los coeficientes cepstrales en la escala Mel (MFCC, *Mel-Frequency cepstral coefficients*) para ambas señales. Las curvas de F0 se representan en la escala de semitonos y se normalizan con respecto a la media para permitir la comparación de las curvas de entonación de diferentes locutores (por ejemplo, voces de hombre y de mujer). A continuación se suavizan los contornos de F0 para eliminar artefactos producidos por la estimación de la frecuencia fundamental. Entonces las dos secuencias de parámetros MFCC se alinean utilizando un alineamiento DTW estándar. Finalmente, las curvas de F0 de las elocuciones de referencia y test se comparan *frame a frame* usando el alineamiento DTW obtenido a partir de las secuencias de observación MFCC. Sin embargo, en lugar de estimar la diferencia entre los patrones de F0 de referencia y de test normalizados *frame a frame*, el presente trabajo propone calcular la correlación entre las dos curvas. Como resultado, las señales de referencia y test se comparan en base a la tendencia descendente-ascendente. La Fig. 3.2 muestra el diagrama de bloques del sistema propuesto de evaluación de acento léxico. En contraste con el método de

evaluación de entonación, el sistema de evaluación del *stress* compara los patrones de referencia y test empleando tanto el F0 como la energía. Como se ha explicado anteriormente, el *stress* es el resultado de la combinación de la intensidad, el pitch y la duración [14]. Si el pitch es la percepción de F0 y el volumen es la percepción de la energía de la señal, entonces tanto el F0 como la energía debieran proporcionar una evaluación más precisa del acento léxico que el F0 o la energía de forma individual.



**Figura 3.1:** Diagrama en bloques del sistema de evaluación de entonación para enseñanza de idiomas.

### 3.3.1. El sistema de evaluación de entonación

#### 3.3.1.1. Pre-procesamiento

En primer lugar, las señales se muestrean a 16 kHz. Luego se aplica un *end-point detector* para eliminar los silencios al comienzo y al final de las señales. Se aplica un

filtro pasa alto con frecuencia de corte 75 Hz para reducir el ruido de baja frecuencia. Finalmente, se aplica un filtro FIR de pre-énfasis  $H(z) = 1 + 0,97z^{-1}$ . Observar que la técnica de alineamiento entre las señales de referencia y test utiliza los coeficientes cepstrales en la escala de Mel. El filtro de pre-énfasis ayuda a ecualizar las diferencias de energía que se observan entre las componentes de baja y alta frecuencia.

### 3.3.1.2. Extracción de F0 y post-procesamiento

Después del pre-procesamiento, las señales son procesadas por un filtro pasa bajos con frecuencia de corte igual a 600 Hz para eliminar aquellas frecuencias fuera del rango de interés. Luego se dividen en *frames* de 400 muestras con 50 % de traslape. La frecuencia fundamental F0 se estima para cada *frame* y se representa en una escala de semitonos:

$$F0_{semitonos}(t) = 12 \frac{\ln[F0(t)]}{\ln(2)} \quad (3.1)$$

donde  $F0(t)$  y  $F0_{semitonos}(t)$  son, respectivamente, la frecuencia fundamental en Hertz y en semitonos para el frame  $t$ . La escala logarítmica permite representar F0 de acuerdo a la percepción humana. Para reducir los errores de *halving* o *doubling* en la estimación de F0, la curva  $F0_{semitonos}(t)$  es suavizada siguiendo un esquema similar al presentado por Zao *et al.* [36] y un filtro mediano. Luego se normaliza respecto a la media. En comparación con la técnica presentada por Peabody y Seneff [66] donde las curvas de F0 son normalizadas respecto a un corpus completo, este trabajo propone una norma-

lización por elocución basada en un esquema *top-down*. Observar que los patrones de entonación en ambas señales de referencia y test son comparadas directamente sin requerir la transcripción o un patrón de F0 predefinido. Finalmente, las discontinuidades causadas por los intervalos áfonos (*unvoiced*) son llenados usando interpolación lineal. La curva de F0 resultante se denota por  $F0_{pp}(t)$ .

### 3.3.1.3. Alineamiento DTW

Se calculan treinta y tres parámetros MFCC por *frame* tanto para la señal acústica de referencia como la de test: la energía más diez coeficientes estáticos y sus derivadas de primer y segundo orden. Luego se utiliza el algoritmo de DTW para alinear las dos secuencias de observación. La distancia local entre los *frames* se estima usando distancia euclidiana o bien la métrica de Mahalanobis. La distancia de Mahalanobis,  $d_{mahalanobis}$ , está dada por:

$$d_{mahalanobis}(O_{t1}^R, O_{t2}^S) = [(O_{t1}^R - O_{t2}^S)^T \Sigma^{-1} (O_{t1}^R - O_{t2}^S)]^{\frac{1}{2}} \quad (3.2)$$

donde  $O_t^R$  y  $O_t^S$  denotan los vectores de observación en el *frame*  $t$  de las elocuciones de referencia y test, respectivamente; y,  $\Sigma$  es la matriz de covarianza de las señales de referencia y de test. En contraste con el enfoque de alineación heurístico propuesto por Delmonte *et al.* [68], el método de programación dinámica presentado en este trabajo es una técnica bien conocida que no requiere reglas, no impone restricciones en el número de características utilizadas en la estimación del alineamiento óptimo y no requiere la

transcripción del texto de la señal acústica de referencia.

La alineación óptima resultante proporcionada por DTW se denota por  $I(k) = \{I_R(k), I_T(k)\}$ ,  $1 \leq k \leq K$ , donde  $I_R(k)$  e  $I_T(k)$  corresponden al los índices de los *frames* de las señales de referencia y test, respectivamente, que son alineadas.

En general, la robustez es un tema clave en procesamiento de voz. En particular, el despliegue masivo de aplicaciones de procesamiento de voz en CALL requiere atenuar el efecto del desacople o *mismatch* de locutor y micrófono. En relación con el *mismatch* de locutor, los distintos niveles de calidad en la pronunciación de segmentos también puede generar una fuente de *mismatch*. La utilización de diferentes tipos de micrófonos de bajo costo es fundamental para el uso masivo de aplicaciones CALL. Por tanto, un conjunto de experimentos presentados en este trabajo pretenden evaluar la robustez del método propuesto, además de su exactitud. De acuerdo a la literatura, es bien sabido que la exactitud de los sistemas de reconocimiento de voz basados en DTW se degrada drásticamente cuando existe *mismatch* de locutor [125, 126, 127] o canal [128]. Sin embargo, el método propuesto en este trabajo utiliza el alineamiento DTW en vez de sus métricas globales como los sistemas de reconocimiento de voz. Como se muestra aquí, las condiciones de *mismatch* de locutor y micrófono tienen un efecto mínimo en la alineación óptima y en la exactitud del sistema.

#### 3.3.1.4. Medida de similitud de F0

A diferencia de la clasificación de F0 como el que presenta Peabody y Seneff [66] para corregir el tono en chino mandarín no nativo, en este trabajo se propone un

sistema de evaluación de la entonación que trata de medir la similitud de la tendencia de la curva de entonación producido por un estudiante y una referencia dada. Observar que en el mandarín hay una serie de tonos léxicos bien definidos [129]. En consecuencia, el problema tratado aquí no es un problema común en clasificación de patrones. Para estimar la medida de similitud de tendencia se comparan las curvas de F0 de referencia y test  $F0_{pp}^R(t)$  y  $F0_{pp}^S(t)$ , respectivamente, *frame a frame* usando el alineamiento DTW. En vez de estimar la distancia acumulada entre  $F0_{pp}^R(t)$  y  $F0_{pp}^S(t)$ , este trabajo propone que ambas curvas sean comparadas desde el punto de vista de su tendencia. En otras palabras, el sistema debiera decidir si el estudiante es capaz de producir una curva de entonación con el mismo patrón que la referencia. Dado el alineamiento DTW entre las señales acústicas de referencia y test,  $I(k)$ , mencionados anteriormente, la medida de similitud entre ambas curvas,  $TS(F0_{pp}^R, F0_{pp}^S)$ , se define como la correlación entre  $F0_{pp}^R(t)$  y  $F0_{pp}^S(t)$ :

$$TS(F0_{pp}^R, F0_{pp}^S) = \frac{\sum_{k=1}^T \{F0_{pp}^R[i_R(k)] - \overline{F0_{pp}^R}\} \{F0_{pp}^S[i_S(k)] - \overline{F0_{pp}^S}\}}{\sigma_{F0_{pp}^R} \cdot \sigma_{F0_{pp}^S}} \quad (3.3)$$

donde  $\sigma_{F0_{pp}^R}$  y  $\sigma_{F0_{pp}^S}$  son las desviaciones estándar de  $F0_{pp}^R(t)$  y  $F0_{pp}^S(t)$ , respectivamente. De forma alternativa, la medida de similitud fue también evaluada usando la distancia euclidiana entre  $F0_{pp}^R(t)$  y  $F0_{pp}^S(t)$ :

$$TS(F0_{pp}^R, F0_{pp}^S) = \sqrt{\sum_{k=1}^T \{F0_{pp}^R[i_R(k)] - F0_{pp}^S[i_S(k)]\}^2}. \quad (3.4)$$

Finalmente, con fines de comparación, se considera también la medida de similitud

entre  $\frac{dF0_{pp}^R[i_R(k)]}{di_R(k)}$  y  $\frac{dF0_{pp}^S[i_S(k)]}{di_S(k)}$  con la correlación y la distancia euclidiana:

$$TS \left( \frac{dF0_{pp}^R[i_R(k)]}{di_R(k)}, \frac{dF0_{pp}^S[i_S(k)]}{di_S(k)} \right) = \frac{\sum_{k=1}^T \left\{ \frac{dF0_{pp}^R[i_R(k)]}{di_R(k)} - \overline{\frac{dF0_{pp}^R}{di_R(k)}} \right\} \left\{ \frac{dF0_{pp}^S[i_S(k)]}{di_S(k)} - \overline{\frac{dF0_{pp}^S}{di_S(k)}} \right\}}{\sigma_{\frac{dF0_{pp}^R}{di_R}} \cdot \sigma_{\frac{dF0_{pp}^S}{di_S}}} \quad (3.5)$$

$$TS \left( \frac{dF0_{pp}^R[i_R(k)]}{di_R(k)}, \frac{dF0_{pp}^S[i_S(k)]}{di_S(k)} \right) = \sqrt{\sum_{k=1}^T \left\{ \frac{dF0_{pp}^R[i_R(k)]}{di_R(k)} - \frac{dF0_{pp}^S[i_S(k)]}{di_S(k)} \right\}^2} \quad (3.6)$$

$$(3.7)$$

donde:

$$\frac{dF0_{pp}^R(i_R)}{di_R} = \begin{cases} F0_{pp}^R(i_R) - F0_{pp}^R(i_R - 1) & \text{if } i_R > 0 \\ F0_{pp}^R(1) & \text{si } i_R = 0 \end{cases} \quad (3.8)$$

$$\frac{dF0_{pp}^S(i_S)}{di_S} = \begin{cases} F0_{pp}^S(i_S) - F0_{pp}^S(i_S - 1) & \text{if } i_S > 0 \\ F0_{pp}^S(1) & \text{si } i_S = 0 \end{cases} \quad (3.9)$$

La motivación de usar la derivada de  $F0_{pp}^R$  y  $F0_{pp}^S$  en vez de la representación estática de las curvas se debe a que la primera puede representar mejor la tendencia de subida y bajada de la frecuencia fundamental que necesita ser evaluada.

El sistema de entonación propuesto en este trabajo apunta a clasificar la entonación

de acuerdo a cuatro patrones que son ampliamente usados en lingüística [14, 13, 12]: ascendente alto (*high rise*, HR); descendente alto (*high fall*, HF); ascendente bajo (*low rise*, LR); y, descendente bajo (*low fall*, LF).

### 3.3.2. El sistema de evaluación de acento léxico

El sistema de evaluación de acento léxico, presentado en la 3.2, es generado a partir del esquema de la 3.1. Se incorpora la extracción de la energía (intensidad) y se combina con la curva de F0 post procesada para decidir si el acento de referencia de una señal dada corresponde o no al de una elocución de test. La energía para un frame  $t$ ,  $E(t)$ , se estima como:

$$E(t) = 10 \log \left[ \sum_{n=1}^N x^2(t+n) \right] \quad (3.10)$$

donde  $x(\cdot)$  denota las muestras de la señal acústica y  $N$  es el ancho del frame (en muestras). Si  $E^R(t)$  y  $E^S(t)$  denotan a la curva de energía de las elocuciones de referencia y test, respectivamente, la medida de similitud de tendencia que incluye a F0 y la energía,  $TS(F0_{pp}^R, E^R, F0_{pp}^S, E^S)$ , se calcula de acuerdo a:

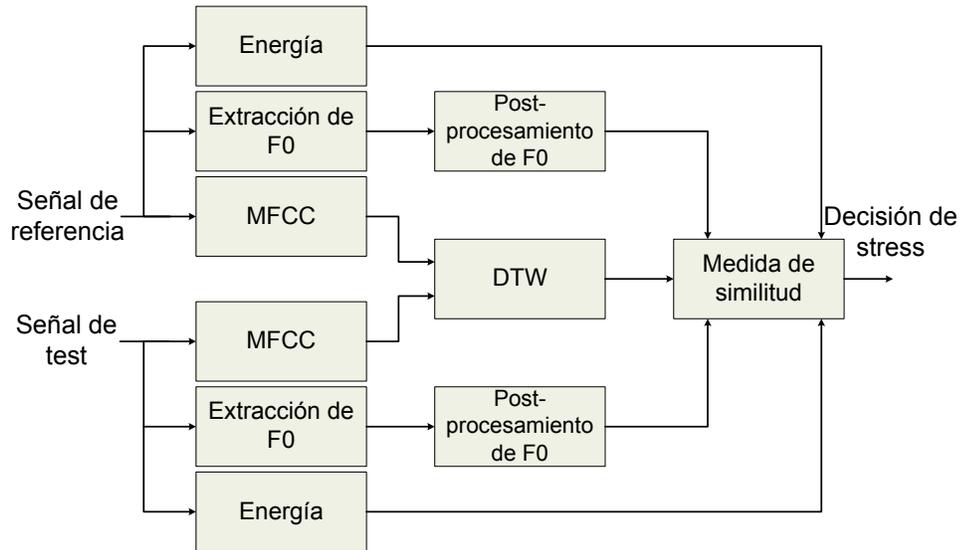
$$TS(F0_{pp}^R, E^R, F0_{pp}^S, E^S) = \alpha \cdot TS(E^R, E^S) + (1 - \alpha) \cdot TS(F0_{pp}^R, F0_{pp}^S) \quad (3.11)$$

donde  $TS(E^R, E^S)$  y  $TS(F0_{post-p}^R, F0_{post-p}^S)$  son estimadas de acuerdo a la ecuación 3.3

haciendo uso de la correlación entre  $E^R$  y  $E^S$ , y entre  $F0_{pp}^R$  y  $F0_{pp}^S$ , respectivamente. El parámetro  $\alpha$  corresponde a un peso que controla la importancia de la energía o la frecuencia fundamental. Finalmente, el sistema toma la decisión sobre el acento léxico de la elocución de test del usuario,  $SD$ , de acuerdo a:

$$SD [TS(F0_{post-p}^R, E^R, F0_{post-p}^S, E^S)] = \begin{cases} \text{igual a la referencia} & \text{si } TS(F0_{post-p}^R, E^R, F0_{post-p}^S, E^S) > \theta_{SD} \\ \text{distinto a la referencia} & \text{en otro caso} \end{cases} \quad (3.12)$$

donde  $\theta_{SD}$  corresponde a un umbral de decisión, el que a su vez depende de la tasa de falsa aceptación o falso rechazo objetivo.



**Figura 3.2:** Diagrama en bloques del sistema de evaluación de acento léxico propuesto.

## 3.4. Experimentos

### 3.4.1. Bases de datos

Dos bases de datos fueron grabadas en el Laboratorio de procesamiento y Transmisión de Voz (LPTV), Universidad de Chile, para evaluar el desempeño de los sistemas propuestos para abordar los problemas de evaluación de entonación y acento léxico en CALL. Todas las señales se grabaron en un entorno de oficina con una frecuencia de muestreo igual a 16 kHz. Existen dos tipos de locutores: expertos y no expertos en idioma inglés y fonética. Los hablantes expertos corresponden a un profesor de idioma inglés, lingüística y literatura inglesa y a sus alumnos de último año del Departamento de Lingüística de la Universidad de Chile. Todos los locutores no expertos demostraron un dominio intermedio de inglés. Tres micrófonos fueron utilizados: un micrófono Shure PG58 (Mic1) y dos micrófonos de PC de escritorio de bajo costo (Mic2 y Mic3). A continuación se describen en detalle las bases de datos.

#### 3.4.1.1. Base de datos para evaluación de entonación

Con el objeto de evitar dificultades adicionales desde el punto de vista del usuario, se consideraron oraciones cortas que no incluyen palabras poco comunes o estructuras sintácticas complejas. Se utilizan los patrones de entonación más comunes: HR, HF, LR, y LF. Observar que en el procedimiento de *testing* se espera que los estudiantes reproduzcan patrones de entonación después oír las oraciones de referencia, contrastando sus realizaciones con la elocución de referencia. El conjunto de datos fue generado a

partir de seis frases en inglés: “*What’s your name*”; “*My name is Peter*”; “*It’s made of wood*”; “*It’s terrible*”; “*It was too expensive*”; and, “*I tried both methods*”. Las oraciones fueron pronunciadas con los patrones de entonación mencionados anteriormente: HR, HF, LF y LR. En total hay 6 oraciones  $\times$  4 patrones de entonación = 24 tipos de elocuciones que fueron grabadas por 16 locutores (ocho expertos y ocho no expertos en idioma inglés y fonética), haciendo uso de tres micrófonos simultáneamente. Entonces, el número total de señales acústicas registradas es igual a 24 tipos de oraciones  $\times$  16 locutores  $\times$  3 micrófonos = 1152 señales. En los experimentos de evaluación de la entonación, las expresiones de referencia corresponden a las frases grabadas por uno de los expertos en idioma inglés y fonética (el más avanzado). El número de experimentos posibles por frase objetivo por locutor por micrófono es igual a 4 etiquetas de patrones de entonación de referencia  $\times$  4 etiquetas de patrones de entonación de test = 16 experimentos. Por último, el número total de experimentos de evaluación de entonación es igual a 16 experimentos por locutor por oración por micrófono por  $\times$  15 locutores de test  $\times$  6 tipos de oraciones  $\times$  3 micrófonos = 4320 experimentos.

#### **3.4.1.2. Base de datos para evaluación de acento léxico**

En primer lugar, por temas de viabilidad, las palabras clave fueron seleccionadas con el fin de evitar el acento secundario. A pesar de que el *stress* secundario es un tema relevante en la adquisición de idiomas ya que puede afectar a la pronunciación de los segmentos, este trabajo se centra en el acento léxico primario. La colocación incorrecta de este acento puede afectar el significado de una palabra dada. En este contexto, las pa-

labras seleccionadas se componen de dos, tres y cuatro sílabas. Este conjunto de datos está compuesto por doce palabras en inglés: “*machine*”; “*alone*”; “*under*”; “*husband*”; “*yesterday*”; “*innocence*”; “*important*”; “*excessive*”; “*melancholy*”; “*caterpillar*”; “*impossible*”; and, “*affirmative*”. Cada palabra fue pronunciada con todas las variantes posibles de acento. El número de variantes posible por palabra es igual al número de sílabas de la misma. Por lo tanto, en total hay  $4 \times$  palabras (2 sílabas + 3 + 4 sílabas sílabas) = 36 tipos de elocuciones grabadas por ocho locutores (cuatro expertos y cuatro no expertos en idioma inglés y fonética), haciendo uso de tres micrófonos al mismo tiempo. Entonces, el número total de grabaciones registradas es igual a 36 tipos de expresiones  $\times$  8 locutores  $\times$  3 micrófonos = 864 grabaciones. En el experimento de evaluación del acento léxico, las oraciones de referencia corresponden a señales grabadas por uno de los expertos en idioma inglés y fonética (el más avanzado). Por último, el número total de experimentos en evaluación de acento es igual a 36 experimentos por locutor por micrófono  $\times$  7 locutores  $\times$  3 micrófonos de prueba = 756 experimentos.

#### **3.4.1.3. Configuración experimental**

El algoritmo DTW mencionado en 3.1 y 3.2 se llevó a cabo de acuerdo con Sakoe y Chiba [38]. La matriz de covarianza usada en la distancia de Mahalanobis en la ecuación 3.2 se estimó con un subconjunto de la base de datos de la evaluación de la entonación se explica en 3.4.1.1. La frecuencia fundamental F0 se calculó mediante el uso de la autocorrelación del sistema Praat [130]. Como se mencionó anteriormente, las expresiones se dividen en *frames* de 400 muestras con traslape igual a 50%. Se

calcularon treinta y tres parámetros MFCC por *frame*: la energía del *frame* más diez coeficientes estáticos y sus derivadas de primer y segundo orden.

#### **3.4.1.4. Puntuación basada en la correlación objetiva-subjetiva**

La correlación subjetiva-objetiva se estima como la correlación entre las puntuaciones subjetivas y objetivas entregados por el sistema de evaluación automática de la entonación propuesto. Las puntuaciones subjetivas se generan de acuerdo al procedimiento que se describe a continuación. En primer lugar, un experto en fonética e idioma inglés (el más avanzado) graba las oraciones con todos los patrones de entonación que se describen en la sección 3.4.1.1. Estas oraciones fueron seleccionados como referencia y cada una de ellas fue etiquetada con HR, HF, LR, o LF (ver sección 3.3.1.4). Luego, los siete expertos restantes escucharon y repitieron cada una de las frases de referencia, siguiendo el patrón de entonación correspondiente. De la misma forma, los ocho no expertos grabaron las oraciones de referencia, pero fueron supervisados por los siete expertos para asegurarse de que el patrón de entonación deseado se generara correctamente. Después, las oraciones grabadas por los siete locutores expertos y los ocho no expertos fueron también etiquetados con HR, HF, LR, o LF. Por último, un ingeniero verificó la concordancia entre las señales acústicas y la etiqueta de entonación asignada. La mayoría de los trabajos en el área del entrenamiento de pronunciación asistido por computador (CAPT, *computer aided pronunciation training*) usan la correlación subjetiva-objetiva como puntuación para evaluar la exactitud de un sistema dado. En este contexto, la Tabla 3.1 define los puntajes subjetivos cuando una elocución de test

pronunciada por un estudiante se compara con una referencia que contiene el patrón de entonación a seguir. En consecuencia, las evaluaciones subjetivas, que resultan de la comparación directa entre las etiquetas de entonación de las señales de referencia y de test, se definen en la Tabla 3.1. Si  $SubjEv_{Testing}$  y  $SubjEv_{Reference}$  denotan la evaluación subjetiva de las oraciones de test y de referencia, respectivamente, donde  $SubjEv_{Testing}$  y  $SubjEv_{Reference}$  son una de las siguientes categorías con respecto al patrón de entonación: HF, LF; HR; y, LR. Por lo tanto, la escala de evaluación subjetiva estricta (Tabla 3.1-a), que resulta de la comparación de la entonación de referencia y test, se define como sigue:

$$\text{Strict subjective score} = \begin{cases} 5 & \text{si } SubjEv_{Testing} = SubjEv_{Reference} \\ 1 & \text{en otro caso.} \end{cases} \quad (3.13)$$

Del mismo modo, la escala no-estricta se define como sigue:

$$\begin{aligned}
& \text{Non-strict subjective score} = \\
& \left\{ \begin{array}{l}
5 \quad \text{si } \text{SubjEv}_{\text{Testing}} = \text{SubjEv}_{\text{Reference}} \\
4 \quad \text{si } (\text{SubjEv}_{\text{Testing}}, \text{SubjEv}_{\text{Reference}}) \in (\text{HF}, \text{LF}), (\text{LF}, \text{HF}), (\text{HR}, \text{LR}), (\text{LR}, \text{HR}) \\
1 \quad \text{en otro caso.}
\end{array} \right.
\end{aligned}
\tag{3.14}$$

Como se muestra en la ecuación 3.14, a las sustituciones HF/LF y HR/LR se asigna un puntaje igual a 4 porque el *score* 3 es neutro y 2 es negativo. Se considera adecuado dar al estudiante un puntaje positivo si él/ella reproduce un patrón de entonación similar a la referencia, aunque no exactamente el mismo.

**Tabla 3.1:** *Escala subjetiva estricta (a) y no-estricta (b) para comparación de entonaciones.*

		(a)						(b)			
		Etiqueta de ref.						Etiqueta de ref.			
Etiqueta	de test	HF	LF	HR	LR	Etiqueta	de test	HF	LF	HR	LR
	<b>HF</b>	5	1	1	1		<b>HF</b>	5	4	1	1
	<b>LF</b>	1	5	1	1		<b>LF</b>	4	5	1	1
	<b>HR</b>	1	1	5	1		<b>HR</b>	1	1	5	4
	<b>LR</b>	1	1	1	5		<b>LR</b>	1	1	4	5

### 3.4.2. Experimentos para medir la exactitud del alineamiento

#### DTW

Como se explicó anteriormente, el *mismatch* de locutor, pronunciación de segmentos y micrófono puede afectar la precisión del alineamiento. Por lo tanto, en este trabajo

se evalúa la exactitud del algoritmo DTW. A partir de la base de datos de entonación (sección 3.4.1.1) se escoge un subconjunto de tres locutores expertos y dos no expertos, para evaluar la robustez del alineamiento DTW. Las señales acústicas de dos micrófonos fueron utilizadas: Shure PG58 y uno de los micrófonos para PC de escritorio de bajo costo. Por lo tanto, un número total de grabaciones es igual a 240. Estas señales fueron segmentadas y etiquetadas fonéticamente en forma manual. El error de alineamiento en la frontera fonética (*phonetic boundary*)  $b$ ,  $E_{align}(b)$  (%), se define como:

$$E_{align}(b) = 100 \cdot \frac{d(b)}{D} \quad (3.15)$$

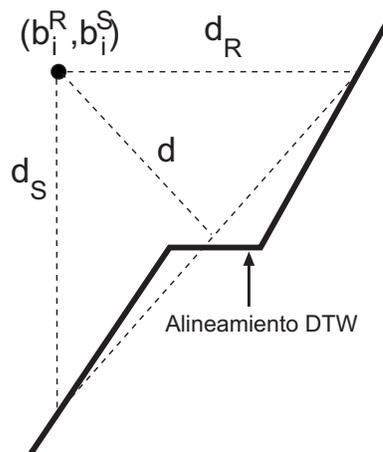
donde  $D$  es el ancho de la ventana de búsqueda en DTW y  $d$  es definido como:

$$d(b) = \frac{1}{2} \sqrt{d_R^2(b) + d_S^2(b)} \quad (3.16)$$

donde  $d_R(b)$  y  $d_S(b)$  son las distancias horizontales y verticales, respectivamente, entre las fronteras fonéticas obtenidas por el etiquetado realizado por humanos y el alineamiento DTW (ver 3.3). Dadas dos señales con la misma transcripción fonética, el error total de alineamiento se estima de acuerdo a:

$$E_{align} = \frac{1}{B} \sum_{b=1}^B E_{align}(b) \quad (3.17)$$

donde  $B$  es el número total de fronteras fonéticas en la elocución.



**Figura 3.3:** Representación de la medida de error,  $d$ , para el alineamiento DTW. El punto  $(b_i^R, b_i^S)$  indica la intersección del límite  $i$  dentro de las señales de referencia y test. Las distancias  $d_R$  y  $d_S$  son las distancias horizontales y verticales, respectivamente, entre los límites fonéticos y el alineamiento DTW.

## 3.5. Resultados y discusiones

### 3.5.1. Experimentos de alineamiento

La Tabla 3.2 muestra el error de alineamiento DTW cuando se usan varias características en combinación con la distancia euclidiana. Se utilizó la base de datos explicada en la sección 3.4.1.1. Todas las señales acústicas que poseen la misma transcripción fueron comparadas de dos en dos de forma independiente del hablante y tipo de micrófono. Como puede verse en la Tabla 3.2, el menor error de alineamiento se obtiene MFCC como característica en combinación la energía de *frame* (estadísticamente significativo con  $p < 0,0001$  cuando se compara con combinaciones de otras características). La Tabla 3.3 compara el error de alineamiento en presencia y ausencia de *mismatch* de locutor, donde se usó distancia euclidiana y Mahalanobis en combinación

con MFCC más energía. Cuando la métrica euclidiana se reemplaza con la distancia de Mahalanobis, el error se reduce en un 10 % (esta diferencia es estadísticamente significativa con  $p < 0,0001$ ). También en la Tabla 3.3 se observa que, cuando se compara con la condición de *mismatch* de locutor, el error de alineamiento muestra un aumento de sólo 1,68 y 1,36 puntos porcentuales cuando las elocuciones fueron generados por distintos hablantes con la distancia euclidiana y Mahalanobis, respectivamente. En consecuencia, este resultado sugiere que el alineamiento DTW es robusto a *mismatch* de locutor.

**Tabla 3.2:** Error de alineamiento DTW usando distintas características. La distancia local corresponde a la métrica Euclidiana.

Característica	Error de alineamiento
Energía	13,27 %
F0	11.49 %
F0 + energía	11.06 %
MFCC	5.31 %
MFCC + energía	4.90 %

La tabla 3.4 muestra error de alineamiento en presencia y ausencia de *mismatch* de micrófono entre las señales acústicas de referencia y de test. Como se puede apreciar, cuando se compara con la condición de ausencia de *mismatch* de micrófono, el error de alineamiento muestra un aumento de sólo 0,12 y 0,03 puntos porcentuales cuando las elocuciones de referencia y test fueron grabadas con el distintos micrófonos, con las distancias euclidiana y Mahalanobis, respectivamente. En consecuencia, a pesar de que la exactitud de los sistemas de reconocimiento de voz basados en DTW se degrada dramáticamente en condiciones de *mismatch*, los resultados mostrados por las

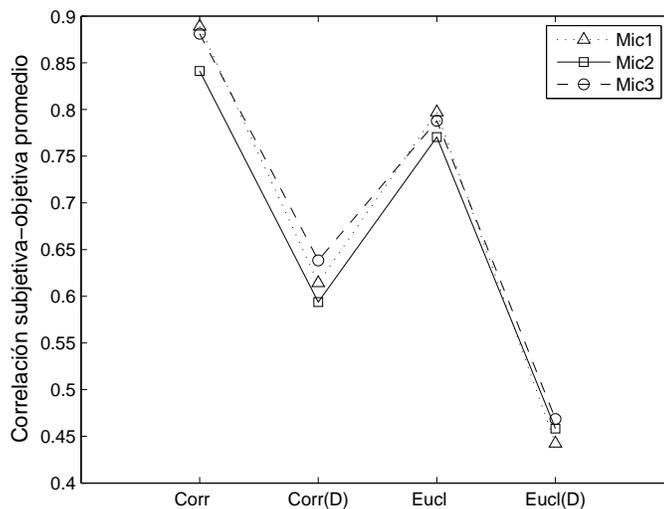
Tablas 3.3 y 3.4 sugieren que el alineamiento DTW es robusto al *mismatch* de locutor y micrófono.

**Tabla 3.3:** *Error de alineamiento DTW con y sin mismatch de locutor.*

Condición referencia-test	Distancia euclidiana	Distancia Mahalanobis
Con mismatch	4,78 %	4,22 %
Sin mismatch	3,10 %	2,86 %

**Tabla 3.4:** *Error de alineamiento DTW con y sin mismatch de micrófono.*

Condición referencia-test	Distancia euclidiana	Distancia Mahalanobis
Con mismatch	3,22 %	2,89 %
Sin mismatch	3,10 %	2,86 %



**Figura 3.4:** *Correlación subjetiva-objetiva promedio en evaluación de entonación para diferentes micrófonos. Mic1 representa el micrófono de alta calidad, mientras que Mic2 y Mic3 corresponden a micrófonos para computador de escritorio de bajo costo.*

### 3.5.2. Experimentos de entonación

La Tabla 3.5 muestra la correlación subjetiva-objetiva promedio entre la medida de similitud de tendencia proporcionada por el sistema de la 3.1 y la evaluación subjetiva de la base de datos de entonación que se menciona en la sección 3.4.1.1. Las evaluaciones subjetivas fueron generadas usando las escalas estricta y no-estricta que se definen en la Tabla 3.1, respectivamente. De acuerdo con la Tabla 3.5, la correlación subjetiva-objetiva promedio más alta está dado cuando se utiliza la correlación como una medida de similitud de tendencia (estadísticamente significativo con  $p < 0,0001$  cuando se compara con las otras medidas de similitud). Al usar la escala de evaluación subjetiva no-estricta, la correlación subjetiva-objetiva promedio es tan alta como 0,88. Sin embargo, con la escala de subjetiva estricta la correlación subjetiva-objetiva disminuye sustancialmente (la disminución es estadísticamente significativa con  $p < 0,0001$ ). Este resultado sugiere que el sistema propuesto es capaz de distinguir con precisión las subidas y las bajabas de la entonación. Por otra parte, la precisión para distinguir entre HF y LF o entre HR y LR es reducida.

**Tabla 3.5:** *Correlación subjetiva-objetiva promedio en evaluación de entonación con distintas medidas de similitud. Las escalas estricta y no-estricta están definidas en la Tabla 3.1.*

Medida de similitud	Escala estricta	Escala no-estricta
Correlación	0,54	0,88
Distancia euclidiana	0,40	0,62
Correlación (D)	0,48	0,79
Distancia euclidiana (D)	0,31	0,46

La robustez de alineamiento DTW en condición de *mismatch* de locutor sugerida por

la Tabla 3.3 se corrobora en la Tabla 3.6, donde se muestra que la correlación subjetiva-objetiva promedio en evaluación de entonación con y sin condición de *mismatch* de locutor. Como se puede observar, la condición de *mismatch* de locutor genera una reducción en la correlación subjetiva-objetiva promedio tan baja como 8,2%. Por otra parte, en el contexto de aprendizaje del segundo idioma, la pronunciación de segmentos fonéticos es también puede ser considerada como una fuente de *mismatch*. La Tabla 3.7 presenta la correlación subjetiva-objetiva promedio en evaluación de entonación con y sin la condición de *mismatch* de pronunciación de segmentos. En el primer caso, las elocuciones de referencia y test provienen de los expertos en idioma inglés y fonética. En el último caso, las señales acústicas fueron pronunciadas por locutores no expertos. De acuerdo con la Tabla 3.7, la pronunciación el *mismatch* de pronunciación de segmentos conduce a una reducción en la correlación subjetiva-objetiva promedio en evaluación de entonación tan baja como 2,5%; 7,6%; 0,5%; y, 0,0% con similitud tendencia estimada con las ecuaciones 3.3, 3.4, 3.5, 3.6, respectivamente. Este resultado sugiere además la validez de la hipótesis sobre la robustez del alineamiento DTW al locutor y a la calidad de pronunciación de segmentos.

**Tabla 3.6:** *Correlación subjetiva-objetiva promedio en evaluación de entonación con distintas medidas de similitud. Se comparan las condiciones con y sin mismatch de locutor, usando la escala no-estricta definida en la Tabla 3.1.*

<b>Medida de similitud</b>	<b>Sin mismatch de locutor</b>	<b>Con mismatch de locutor</b>
Correlación	0,88	0,88
Distancia euclidiana	0,71	0,62
Correlación (D)	0,79	0,79
Distancia euclidiana (D)	0,57	0,46

**Tabla 3.7:** *Correlación subjetiva-objetiva promedio en evaluación de entonación con distintas medidas de similitud. Se comparan las condiciones con y sin mismatch de pronunciación de segmentos, usando la escala no-estricta definida en la Tabla 3.1.*

Medida de similitud	Sin mismatch de locutor	Con mismatch de locutor
Correlación	0,89	0,87
Distancia euclidiana	0,65	0,60
Correlación (D)	0,79	0,79
Distancia euclidiana (D)	0,44	0,53

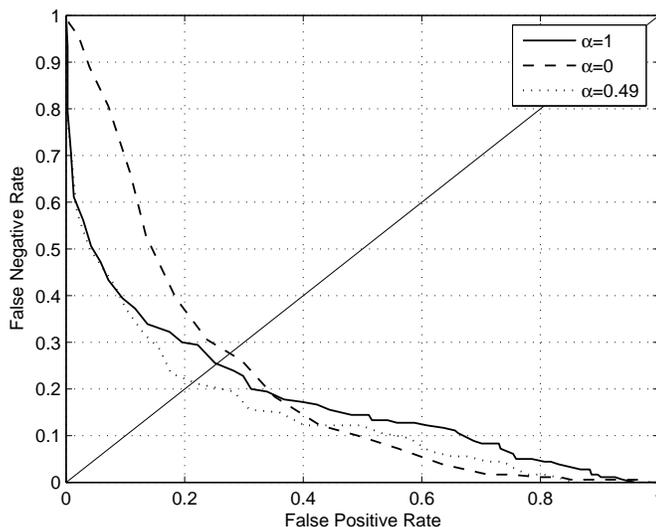
La 3.4 muestra la correlación subjetiva-objetiva promedio en evaluación de entonación con y sin *mismatch* de micrófono. Las señales acústicas de referencia fueron grabadas con Mic1. Las elocuciones de test fueron capturadas con los micrófonos Mic1, Mic2 y Mic3. Como se puede ver en la 3.4, la diferencia en la correlación subjetiva-objetiva promedio en evaluación de la entonación entre las condiciones con y sin *mismatch* es, en promedio, igual a 2,5%. Este resultado corrobora aquel discutido en la Tabla 3.4.

### 3.5.3. Experimentos de acento léxico

La Fig. 3.5 muestra las curvas ROC (*receiver operating characteristic*), que se obtiene graficando la tasa de falsos negativos, FNR (*false negative rate*), y la tasa de falsos positivos, FPR (*false positive rate*), con el sistema de evaluación del acento léxico que es presentado en este trabajo (Fig. 3.2). La medida de similitud de tendencia se calcula con la ecuación 3.11 y la decisión final sobre la evaluación del acento se toma de acuerdo a la ecuación 3.12. La variable  $\alpha$  se ajusta con el fin de minimizar el área bajo la curva ROC. El valor óptimo obtenido es igual a 0,49. La Fig. 3.5 también muestra las curvas de FPR/FNR con  $\alpha = 0$ ,  $\alpha = 1$ , y  $\alpha = 0,49$ . La Tabla 3.8 presenta el área bajo

la curva ROC y el EER (*equal error rate*), con  $\alpha$  igual a 0, 1 y 0,49. De acuerdo con la Fig. 3.5 y la Tabla 3.8, el  $\alpha$  óptimo entrega una reducción en el área bajo la curva ROC y en el EER igual a 15,5% y 22,3%, respectivamente, cuando se compara con  $\alpha = 1$  y  $\alpha = 0$ . Usando el test de significancia de McNemar [131], se concluye que las diferencias en EER entre  $\alpha = 0,49$ , y entre  $\alpha = 0,49$  y  $\alpha = 0$  son estadísticamente significativas con  $p < 0,00048$  y  $p < 0,077$ , respectivamente. Este resultado sugiere que tanto la frecuencia fundamental como la energía proporcionan información relevante para evaluar el acento léxico de una palabra dada. La exactitud del sistema de evaluación de *stress* debe podría mejorar incluyendo la información de duración, lo que a su vez no es fácil de realizar en el marco del alineamiento DTW. Sin embargo, vale la pena mencionar que los sistemas de evaluación de calidad fonética a nivel de palabras en el estado del arte entregan correlaciones subjetiva-objetiva entre 0,6 y 0,8, dependiendo, entre otros factores, del número de niveles en la escala de evaluación [54, 55, 56, 57, 58, 59, 60, 61]. De acuerdo a Molina *et al.* [53], el error de clasificación, que se define como la diferencia entre las evaluaciones subjetivas y objetivas, fue estimado en un sistema CAPT a nivel de palabras en escalas con dos y cinco niveles. Con la escala de dos niveles, la correlación subjetiva-objetiva es igual a 0,8 en promedio y el error de clasificación está en torno al 10%. Con una escala de cinco niveles, la correlación subjetiva-objetiva es de 0,67 en promedio y el error de clasificación es del orden de 55%. Como resultado, el EER óptimo proporcionado por la técnica de evaluación de acento léxico presentada en este trabajo (21,5%) es similar al que se obtiene con los sistemas de evaluación de

pronunciación a nivel de segmentos. Esto sugiere que el sistema propuesto debiera ser lo suficientemente exacto para aplicaciones reales.



**Figura 3.5:** Curva ROC (false negative versus false positive en evaluación de acento léxico. La medida de similitud fue calculada de acuerdo a la ecuación 3.11 y la decisión se estimó usando la ecuación 3.12. El valor  $\alpha = 1$  indica que se utiliza sólo el contorno de F0 y  $\alpha = 0$  indica que solamente la energía.

**Tabla 3.8:** Área bajo la curva ROC y equal error rate (EER) para evaluación de acento léxico para distintos valores de  $\alpha$  usando correlación como medida de similitud. El  $\alpha$  óptimo que minimiza el EER es 0,49.

Característica	Área ROC	EER (%)
$\alpha = 1$	0,181	25,41
$\alpha = 0$	0,212	27,64
$\alpha = 0,49$	0,147	21,48

El método propuesto requiere un patrón de entonación de referencia que el estudiante debe tratar de seguir. Sin embargo, no se necesita la transcripción de la referencia ni de la señal de test. La motivación detrás de la estrategia propuesta, como se explica aquí, es el hecho de que no hay una definición clara de entonación “correcta”

o “incorrecta” [109]. La misma frase puede ser pronunciada con varios patrones de entonación diferentes de acuerdo al contexto y en la mayoría de los casos existe más de una entonación correcta. El problema abordado en este trabajo consiste en cómo enseñar a un estudiante a seguir un patrón de entonación de referencia dado que no existe sólo una producción de entonación correcta. En contraste, desarrollar un sistema de anotación para entonación está fuera del alcance de la hipótesis considerada en el presente trabajo. Dado este contexto, esta tesis no considera evaluar la entonación en enseñanza de idiomas sin una referencia dada.

### 3.6. Conclusiones

En este capítulo se presentó una discusión sobre la naturaleza y la importancia de la entonación en el aprendizaje de un segundo idioma. Como consecuencia, se propuso un sistema automático para evaluar la entonación basado en un esquema *top-down*. El sistema es independiente del texto y del idioma. Adicionalmente, se presentó un sistema de evaluación del acento léxico que combina la información de la frecuencia fundamental y la energía. El sistema compara directamente la oración pronunciada por el estudiante con una elocución de referencia. La medida de similitud de tendencia de entonación y de energía se comparan *frame a frame* mediante el uso de alineamiento DTW. Además, se aborda el problema de la robustez del alineamiento al locutor, micrófono, calidad de pronunciación. El sistema de evaluación de la entonación alcanza una correlación subjetiva-objetiva promedio tan alta como 0,88 cuando se utiliza la

correlación como medida de similitud de tendencia. Por su parte, el sistema evaluación de acento logra un EER igual a 21,5%, que a su vez es similar al observado en los sistemas de evaluación de pronunciación de segmentos. Estos resultados sugieren que los sistemas propuestos podrían ser utilizados en aplicaciones reales. A pesar de que el sistema fue probado en el marco de aprendizaje de inglés como segundo idioma con hablantes nativos del español, el método propuesto es aplicable a cualquier idioma.

# Capítulo 4

## Modelación prosódica para detección de emociones usando modelos de referencia

En virtud de los resultados obtenidos en el capítulo 3 para CALL, se propone el uso de modelos de referencia para detectar la prominencia emocional, de forma localizada, en la frecuencia fundamental F0. En primer lugar, se considera el caso ideal donde una oración de referencia neutra con la información léxica misma que la elocución de prueba está disponible. Una vez que las señales acústicas de referencia y test son alineadas en el tiempo mediante programación dinámica (DTW, *dynamic time warping*), las curvas de F0 son extraídas de ambas elocuciones y luego se comparan directamente. Los resultados muestran que una sola señal de referencia puede ser utilizada para capturar la

modulación emocional transmitida en F0. Después, el análisis se extiende para modelar la variabilidad intrínseca de la frecuencia fundamental. En vez de utilizar sólo una curva de F0 como *template*, se generan modelos de referencia utilizando una familia de curvas. Para ello, se presenta un nuevo enfoque basado en *functional data analysis* (FDA), que se realiza con modelos tanto dependientes como independientes del léxico. Los modelos neutros se representan mediante una base de funciones y la curva F0 de test es caracterizada por las proyecciones sobre dicha base. Los resultados experimentales muestran que el sistema propuesto permite obtener una exactitud tan alta como 75,8% en clasificación de emociones binaria (i.e. neutro versus emocional), la que a su vez es 6,2% superior a la exactitud alcanzada por un sistema estándar. El análisis se extiende a nivel de sub-oración para detectar los segmentos que son emocionalmente más relevantes. El enfoque se valida mediante una base de datos natural. Los resultados indican que el sistema propuesto puede ser utilizado eficazmente en aplicaciones reales de detección emociones en señales de voz.

## 4.1. Introducción

La comprensión emocional es una habilidad crucial en la comunicación humana, ya que juega un rol preponderante no solamente en las interacciones interpersonales, sino que también en muchas otras actividades cognitivas como la toma racional de decisiones, la percepción y el aprendizaje [3]. Por esta razón, la modelación y el reconocimiento de emociones es esencial en el diseño e implementación de interfaces

hombre-máquina (HMI) que están más en sintonía con las necesidades del usuario. La información emocional se transmite a través de la expresión facial [74], los gestos [75] y la voz [76, 77, 78]. Entre las características acústicas de la voz, la prosodia es uno de los aspectos más importantes. Los cambios en la entonación, el volumen y la duración son usados por las personas para expresar emociones. Como resultado, las características extraídas de la frecuencia fundamental F0, la energía y la duración (es decir, las correlaciones acústicas de prosodia) han sido ampliamente usadas en la literatura para estudiar la modulación emocional en la voz [77, 132]. El estado del arte en detección y reconocimiento de emociones consiste en calcular un conjunto de estadísticas globales o funcionales, como media, la varianza, el rango, el máximo y el mínimo de los descriptores de bajo nivel (por ejemplo, el contorno de F0 y la energía). Después, se aplican algoritmos de selección de características para elegir un subconjunto con los parámetros más relevantes emocionalmente [133]. Este enfoque asume que todos los *frames* en la señal acústica son igualmente importantes. Sin embargo, diversos estudios han demostrado que la información emocional no se distribuye uniformemente en el tiempo [92, 96]. Por ejemplo, la entonación en una elocución alegre muestra una tendencia ascendente al final de la oración [134]. Sin embargo, dado que las estadísticas se calculan a nivel global, no es posible identificar los segmentos más relevantes de forma local dentro de la oración. La detección de estos segmentos emocionalmente más sobresalientes puede conducir al desarrollo de algoritmos de reconocimiento de emociones basados en el proceso de exteriorización de los rasgos emocionales.

Busso *et al.* ha introducido la idea de detectar emociones mediante el uso de modelos neutros (no emocionales) [95, 77]. La hipótesis detrás de este enfoque es que los patrones de la voz emocional difiere de los patrones observados en el habla neutra y, por lo tanto, estas diferencias pueden ser cuantificadas en el espacio de características. La principal ventaja de este esquema es que la disponibilidad de bases de datos neutrales corpus es mayor que en el caso de las bases de datos emocionales, por lo tanto, es posible construir modelos robustos e independientes del locutor. La elocución de test se contrasta con los modelos de referencia generados usando señales neutrales. La verosimilitud de los modelos se utiliza como una medida de *fitness* para caracterizar la señal, ya sea como emocional o neutra. Este enfoque ha sido implementado usando estadísticas globales extraídas de parámetros prosódicos [77] y características espectrales [95]. Este trabajo se basa en las ideas mencionadas anteriormente para detectar la prominencia emocional en forma localizada en el contorno de F0. El método propuesto genera un perfil o *template* de referencia emocionalmente neutro para la curva de F0 el que se compara con la señal acústica de test.

En primer lugar, se evalúa el esquema basado en la comparación de un contorno de F0 con una referencia. La referencia corresponde a una curva de F0 extraída de una elocución neutra con la misma información léxica que la señal de test. Los coeficientes cepstrales en la escala de Mel (MFCC, *Mel-frequency cepstral coefficients*) se extraen de ambas elocuciones. A continuación, las secuencias de MFCC se alinean usando DTW (*dynamic time warping*) a fin de comparar las contornos de F0 *frame* a

*frame*. Finalmente, la se estima la similitud entre las curvas de F0 de referencia y test mediante la correlación de Pearson. Los resultados sugieren que la comparación entre elocuciones de referencia emocionalmente neutras con señales de test puede utilizarse para discriminar entre voz emocional y neutro.

Dado que un patrón de referencia único no es suficientemente representativo de la variabilidad inter e intra-locutor, este trabajo también propone una técnica para generar perfiles de referencia usando un conjunto de curvas emocionalmente neutras provenientes de varios locutores. Para ello, se propone un método basado en *functional data analysis* (FDA) que permite generar una base de funciones a partir de contornos de F0 neutros extraídos de señales que poseen el mismo contenido léxico entre sí. Entonces, la curva de F0 de test se proyecta sobre la base funciones de referencia. Como resultado, las proyecciones del contorno F0 sobre las bases funciones genera un conjunto de parámetros que son utilizados para discriminar entre voz neutra y emocional. El método propuesto alcanza exactitudes tan altas como 75,8% en una reconocimiento binario de emociones (detección de emociones), esto es, voz neutra versus emocional. A su vez, la técnica propuesta entrega una precisión que es 6,2% mayor que aquella alcanzada por un sistema estándar entrenado con estadísticas derivadas de F0. El mismo criterio se evalúa con modelos independientes del léxico que fueron construidos con elocuciones neutras con contenido léxico diferente. Los resultados sugieren que la reducción en la exactitud no es significativa (esto es, de 75,8% a 74,2%) cuando los *templates* de referencia dependientes del léxico se sustituyen por modelos independien-

tes del léxico.

Por último, el enfoque propuesto se aplica a nivel de sub-oración con el propósito de encontrar las partes más sobresalientes desde el punto de vista emocional dentro de una sentencia dada. Para ello, las señales acústicas se dividen en distintos tipos de segmento: en palabras, frases y ventanas de duración fija. Para cada tipo de segmento, se genera un perfil de referencia texto independiente usando el esquema basado en FDA antes mencionado. Dada una elocución de test, se estiman las proyecciones de los segmentos en la base de funciones de referencia con el objeto de determinar las secciones más sobresalientes desde el punto de vista emocional dentro de la oración. Cada segmento se clasifica como neutro o emocional, y se utiliza el promedio del *score* de clasificación para determinar si la oración completa es neutra o emocional. Los resultados muestran que para clasificación neutro/emocional, la exactitud alcanzada cuando se emplean modelos de referencia texto dependientes es similar cuando se compara con la exactitud obtenida usando modelos independientes del léxico. El sistema se valida a nivel de segmentos sub-oración con la base de datos espontánea (i.e. no actuada) SEMAINE [135]. La correlación entre la derivada de las evaluaciones subjetivas y el puntaje entregado por el método propuesto es tan alta como 0,44 (con *inter-evaluator agreement* = 0,36). Este resultado sugiere que el sistema es capaz de detectar los segmentos más prominentes desde el punto de vista emocional dentro de una oración.

Los contornos de F0 se usaron en el capítulo 3 para evaluar la entonación en enseñanza de segundo idioma (ver también [136]). Sorprendentemente, el problema de

reconocimiento de emociones en señales de voz no ha sido abordado usando *templates* que modelen la curva de F0. Dentro de las contribuciones del presente capítulo se tiene: (a) un marco novedoso para la detección de la modulación emocional basado en perfiles de referencia que modela los contornos de F0 generados con voz neutra; (b) un profundo y exhaustivo análisis de las referencias neutras como un método para detectar las emociones en señales de voz; (c) la generación de *templates* de referencia de contorno F0 con *functional data analysis*; y, (d) un estudio de la unidad de segmentación más corta que se puede utilizar en detección de emoción. Cabe destacar que, como se sugiere en [77], el uso de señales acústicas neutras para generar modelos de referencia reduce significativamente la dependencia de bases de datos emocionales (actuadas o espontáneas), las que a su vez son mucho más difíciles de conseguir en comparación con un corpus ordinario. Esto es muy interesante desde el punto de vista de investigación y de aplicación.

Este capítulo se organiza de la siguiente manera: en la sección 4.2 se describe el trabajo relacionado y se proporciona el contexto en el que se desarrolla la contribución de este trabajo. La sección 4.3 evalúa la viabilidad de usar solamente un contorno de F0 extraído de una señal acústica emocionalmente neutra como perfil de referencia para detectar la emoción. La sección 4.4 presenta la estimación de modelos de referencia utilizando una familia de curvas de F0 mediante FDA a nivel de oración. A continuación, la sección 4.5 analiza el enfoque propuesto a nivel de sub-oración (es decir, a nivel de frase, palabra o ventana de duración fija) y muestra la evaluación de la técnica

propuesta usando un corpus emocional espontáneo. Finalmente, la sección 4.6 presenta la discusión, conclusiones y direcciones del trabajo futuro.

## 4.2. Antecedentes

### 4.2.1. Trabajo relacionado

El estudio de la emoción humana ha ganado la atención de varias disciplinas, entre las cuales se tiene la lingüística, la psicología, las ciencias de la computación y la ingeniería. En consecuencia, se ha registrado un número creciente de publicaciones acerca de los avances en reconocimiento y detección automática de emociones. Cowie *et al.* [4], Zeng *et al.* [137] y Schuller *et al.* [138] muestran una completa revisión de los trabajos más importantes. El estado del arte en el reconocimiento de las emociones consiste en la estimación de estadísticas globales de descriptores de bajo nivel como el F0, la energía y los MFCCs. Normalmente, también se incorporan características léxicas. Entre las características prosódicas, las estadísticas globales como la media, el máximo, el mínimo y el rango son considerados como los parámetros emocionalmente más prominentes [77].

Una de las limitaciones de las estadísticas globales es el hecho de que no capturan las variaciones locales observadas en los contornos de F0, las que a su vez podrían proporcionar información útil para la detectar emociones. Patterson y Ladd argumentaron que el rango (esto es, la diferencia entre el máximo y el mínimo de contorno

de F0 de una elocución) no entrega información sobre la distribución de F0 y por lo tanto, se pierde información emocional valiosa [86]. De acuerdo con Lieberman y Michaels [87], pequeñas variaciones bajas en F0 pueden ser subjetivamente relevantes en la identificación de las emociones.

En la literatura, ha habido algunos autores que han intentado modelar la forma del contorno de F0. Paeschke y Sendlmeier [139] analizaron los movimientos de subida y bajada de F0 en los acentos en el habla afectiva. El estudio incorpora métricas relacionadas con los *peaks* de acento dentro de una oración. Los autores encontraron que tales métricas presentan diferencias estadísticamente significativas entre clases emocionales. Además, Paeschke modeló la tendencia global del contorno de F0 en el habla emocional como la pendiente de la regresión lineal [88]. El autor concluyó que la tendencia global puede ser útil para describir emociones como aburrimiento y tristeza. Rotaru y Litman utilizaron los coeficientes de regresión lineal y cuadrática además del error de regresión como características para representar las curvas de F0 [90]. Yang y Campbell argumentaron que la concavidad y convexidad del contorno de F0 reflejan el estado expresivo subyacente [89].

El sistema ToBI (*Tone and Break Indices*) es un sistema de etiquetado de prosodia que ha sido ampliamente utilizado para transcribir la entonación [140]. Liscombe *et al.* analizaron el habla emocional en características acústicas haciendo uso de las etiquetas ToBI para identificar el tipo de acento de pitch nuclear, el tipo de contorno de pitch y los límites de oración [141]. A pesar de que ToBI ofrece un enfoque interesante

para describir los contornos de F0, se requiere un etiquetado más preciso para generar transcripciones prosódicas. En este contexto, Taylor introdujo el modelo *Tilt Intonation* [142] para representar la entonación como una secuencia lineal de los acontecimientos (por ejemplo, acentos tonales o límites), que a su vez son dadas por un conjunto de parámetros. Sin embargo, se requiere un algoritmo de segmentación automática de eventos para utilizar este sistema y, por tanto, no es fácilmente aplicable en las tareas de tareas de reconocimiento o detección de emociones.

A pesar de los esfuerzos por abordar el problema de la caracterización de la voz emocional por medio del modelamiento del contorno F0, esto sigue siendo una tarea abierta. Este trabajo propone un enfoque novedoso basado en una referencia neutral o una *template* para contrastar con una curva de F0 extraída de una elocución de test *frame a frame*. El resultado es una técnica que relaja la restricción de la disponibilidad de bases de datos emocionales y hace posible la detección de los segmentos emocionalmente más relevantes dentro de una oración. El esquema que se presenta también puede extenderse a otras características como la energía, los parámetros espectrales e incluso características que no son voz como por ejemplo los descriptores faciales. Se debe tener en cuenta que el objetivo es discriminar entre voz neutra y emocional (es decir, clasificación binaria). Este problema es más general que un sistema de clasificación multiclase adaptada a un dominio en particular. Un sistema de detección de las emociones puede ser utilizado a través de dominios diferentes, independientemente de las etiquetas emocionales requeridas, atributivos o categorías impuestas por la aplica-

ción de destino. Además, se puede utilizar como un primera etapa en un sistema de reconocimiento de emociones multiclase más sofisticado, en el que las muestras de voz se asignan a etiquetas emocionales más finas (por ejemplo, felicidad o ira).

#### 4.2.2. Bases de datos emocionales

Este capítulo considera tres bases de datos emocionales (ver Tabla 4.1). Estas bases de datos ofrecen las condiciones controladas requeridas por los experimentos propuestos. Dos de estas bases de datos fueron grabadas por actores. A pesar de que las emociones actuadas difieren de aquellas manifestadas en la vida real, el habla afectiva expresada por actores es considerada una buena primera aproximación. La tercera base de datos corresponde a un corpus espontáneo que se utiliza para validar el enfoque propuesto. El análisis presentado en la sección 4.3 se desarrolla en condiciones controladas con experimentos dependientes del léxico y del locutor. Para ello, se requiere que una oración dada sea pronunciada varias veces por un mismo locutor con diferentes estados emocionales, incluyendo el estado neutral. Por esta razón, consideramos la base de datos EMA grabada en la University of Southern California (USC) <sup>1</sup> [143]. Un hombre (*ab*) y dos mujeres (*jn*, *ls*) participaron en la grabación (dos de ellos con entrenamiento teatral formal). Ellos leyeron diez oraciones en inglés con los estados emocionales felicidad, ira, tristeza y el estado neutral (10 oraciones  $\times$  5 repeticiones  $\times$  4  $\times$  3 emociones = 600 muestras en total – *ab* leyó 4 frases adicionales las que a su vez generaron 80 muestras adicionales). Algunos ejemplos de las oraciones son “*I hear the echo of voices*

---

<sup>1</sup>La base de datos EMA está disponible en [http://sail.usc.edu/ema\\_web](http://sail.usc.edu/ema_web)

*and the sound of shoes*” y *“They think the company and I will have a long future”*. A los locutores se les pidió que grabaran las frases en orden aleatorio para atenuar o eliminar reproducciones con entonación similar. Para reducir la fatiga, la grabación se dividió en sesiones pequeñas separadas por descansos. Esta base de datos también contiene información articulatoria, la que no es considerada en este trabajo. La base de datos fue grabada a 16 kHz. El corpus EMA fue evaluado por cuatro hablantes nativos de inglés americano. Los evaluadores seleccionaron las etiquetas emocionales que mejor representan a las señales acústicas de acuerdo a las clases feliz, enojado, triste, neutral y otra. La tasa promedio de reconocimiento humano fue de 81,8 % [144].

A partir de la sección 4.4.1, el análisis no requiere experimentos dependientes del locutor. Como el requisito de tener frases pronunciadas varias veces por el mismo locutor no es necesaria, se considera la base de datos emocional Berlín (EMO-DB) [145]. Esta base de datos se compone de diez locutores (cinco hombres y 5 mujeres), quienes leyeron diez oraciones distintas en alemán, una vez cada una, expresando seis emociones diferentes (miedo, asco, alegría, aburrimiento, tristeza e ira), además del estado neutral. Esta base de datos ha sido ampliamente utilizada en trabajos relacionados con reconocimiento de emociones.

En las últimas secciones del presente capítulo, el *framework* propuesto se extiende relajando el requerimiento de dependencia del léxico. En este caso, se evalúa la exactitud del sistema mediante un corpus emocional espontáneo. El estudio considera la base de datos SEMAINE, que incluye las grabaciones audiovisuales de interaccio-

nes hombre-máquina naturales [135]. Las emociones son provocadas usando el enfoque *sensitive artificial listener* (SAL). En este trabajo se consideran sesiones grabadas por diez locutores. Los datos de SEMAINE contienen además evaluaciones subjetivas generadas por los humanos utilizando el sistema Feeltrace [97]. Esta es una herramienta utilizada para realizar un seguimiento continuo del estado emocional percibido en el tiempo (opuesto a la asignación de una etiqueta discreta por frase). A los evaluadores se les pide para mover el cursor mientras ven y escuchan un estímulo mediante una interfaz gráfica (GUI, *graphical user interface*). La interfaz gráfica registra la posición del puntero, que a su vez describe el contenido emocional en términos de atributos continuos. A pesar de que la base de datos ha sido etiquetada usando diversos atributos emocionales, en este trabajo se considera sólo las dimensiones activación/excitación (calmo versus activo) y valencia (negativo versus positivo) (ver sección 4.5).

### 4.2.3. Extracción de F0 y post-procesamiento

La frecuencia fundamental se calcula mediante un procedimiento equivalente al presentado en el capítulo 3. Primero, las señales acústicas se dividen en *frames* de 400 muestras (25 milisegundos), con traslape de 50 %. La frecuencia fundamental se calcula mediante el uso del sistema de detección de F0 basado en autocorrelación Praat [130]. Después, el F0 de cada *frame* se representa de acuerdo a una escala de semitonos:

$$F0_{semitone}(t) = 12 \cdot \frac{\log[F0(t)]}{\log(2)} \quad (4.1)$$

donde  $F0(t)$  y  $F0_{semitone}(t)$  son la frecuencia fundamental para el *frame*  $t$  en Hertz y semitonos, respectivamente. El esquema propuesto en este trabajo tiene como objetivo modelar el contorno F0 para comparar la voz neutra con la emocional. En este sentido, el logaritmo intenta representar las diferencias de F0 de acuerdo a una escala de percepción semejante a la humana. Después de estimar  $F0_{semitone}(t)$ , los segmentos áfonos son interpolados usando una spline cúbica para obtener contornos de F0 suaves y continuos. Finalmente, la curva resultante  $F0_{semitone}(t)$  se normaliza de restando su media. A partir de ahora, el término “contorno de F0” denota la curva de F0 en semitonos, interpolada y normalizada por la media.

### **4.3. Análisis de la prominencia emocional usando una única señal como referencia**

El propósito de esta sección es mostrar que un señal acústica de referencia neutra puede ser usada para ser contrastada con voz emocional. El experimento que aquí se presenta trata de comparar directamente el contorno de F0 extraído de la señal de test y las elocuciones de referencia neutras que contienen la misma información léxica. La comparación consiste en estimar una medida de similitud entre el contorno de F0 de test y el de referencia. Esta medida de similitud se utiliza para caracterizar a la señal de test como neutral o emocional. Notar que este caso corresponde a un escenario ideal donde tanto las elocuciones de test como de referencia proporcionan la misma

**Tabla 4.1:** Descripción de las bases de datos.

Corpus	Tipo	Uso de los datos	Espontánea/Actuada	# locutores	# señales	Emociones/Atributos
WSJI	Neutral	Referencia	Espontánea	50	8104	<i>neutral</i>
EMA	Emocional	Entrenamiento/Test	Actuada	3	680	<i>neutral, anger, happiness, sadness</i>
EMO-DB	Emocional	Entrenamiento/Test	Actuada	10	535	<i>neutral, fear, disgust, happiness, boredom, sadness, anger</i>
SEMAINE	Emocional	Test	Espontánea	10	–	<i>valence, activation/arousal, power, anticipation/expectation, intensity</i>

información léxica. En las secciones 4.4 y 4.5 se extiende y generaliza este enfoque a escenarios menos restrictivos.

La medida de similitud se estima mediante una estrategia similar a aquella mostrada en el capítulo 3, donde dos señales (es decir, de referencia y test) se comparan en un esquema *top-down* [136]. A fin de mantener todas las variables bajo control exceptuando la modulación emocional, el análisis se realiza en un esquema dependiente del locutor. Además, se consideran elocuciones de referencia y test con el mismo contenido léxico. Teniendo en cuenta estas limitaciones, el análisis mostrado a continuación se lleva a cabo utilizando la base de datos de EMA. A pesar que algunos investigadores han propuesto el uso de unidades de voz más cortas para el análisis de las emociones [146, 147], en esta sección se considera a la oración como unidad de segmentación, ya que unidades más cortas pueden degradar información suprasegmental importante transmitida en F0 [77].

En primer lugar, se extrae la frecuencia fundamental tanto de la elocución de referencia como de test y se aplica el post-procesamiento detallado en la sección 4.2.3. A continuación, las señales de referencia y test se alinean de acuerdo a sus MFCCs mediante el uso de la técnica DTW (se utiliza la distancia euclidiana como métrica, se escoge la condición  $P = 0$  como restricción local y la banda Sakoe-Chiba como restricción global [38]). Por último, se utiliza la correlación de Pearson como medida de similitud para estimar las diferencias entre ambos patrones de F0. Los niveles más bajos de correlación indicarán mayores diferencias entre las frases neutras y emocionales,

los que pueden ser asociados principalmente a la modulación emocional la elocución de test.

Dado un locutor, cada frase neutra es comparada con sus versiones emocionales (es decir, feliz, enojado y triste). Hay 20 realizaciones por oración y por locutor. Cinco de ellos corresponden a voz neutra y 15, a voz emocional. Por lo tanto, el número comparaciones neutra-emocionales posibles por oración y locutor es igual a 75 (5 neutral  $\times$  15 emocional). Esto da un total de 2250 experimentos (75 experimentos  $\times$  10 oraciones  $\times$  3 locutores). Asimismo, se realizó la comparación de señales de test de referencia neutras. El número experimentos neutro-neutro es igual a 10 pares de elocuciones neutras por oración y locutor. Esto da un total de 300 experimentos (10 experimentos  $\times$  10 oraciones  $\times$  3 locutores).

La figura 1 presenta la distribución de la medida de similitud basada en la correlación de las señales de test y referencia para cada emoción. De acuerdo con la Fig. 1, la comparación de elocuciones de test neutras con patrones de referencia (también neutros por definición) muestra correlaciones más altas ( $\rho = 0,84 \pm 0,15$ ). Este resultado muestra que las elocuciones emocionalmente neutras con el mismo contenido léxico y pronunciadas por el mismo locutor producen contornos de F0 similares. Por el contrario, la similitud entre los contornos de F0 proporcionados por elocuciones de test emocionales y referencias neutras es significativamente menor. Notar que la desviación estándar de la medida de similitud para comparaciones neutro-neutro y neutra-triste es menor que aquella obtenida para comparaciones neutro-felicidad y neutra-enojo. La

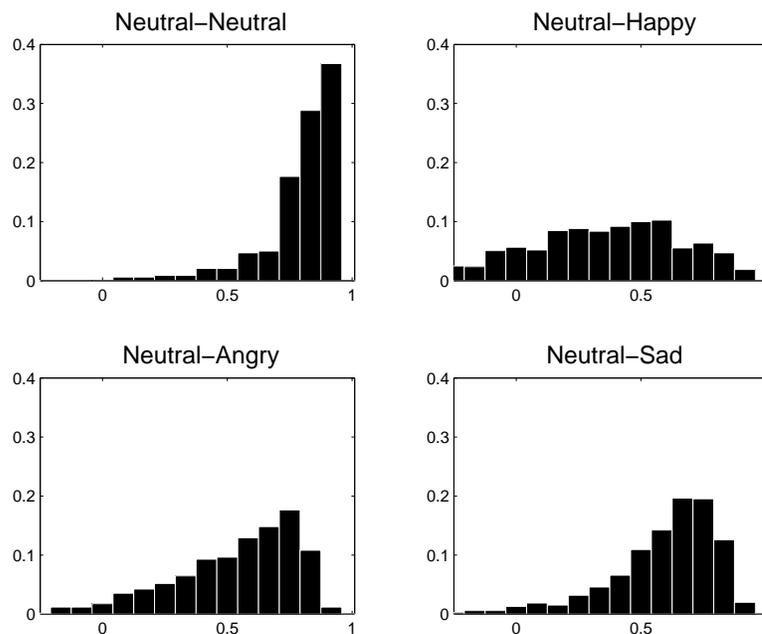
Figura 1 también sugiere que el histograma de la correlación de neutro-felicidad frases presenta la más fuerte divergencia comparado con el histograma neutro-neutro. En consecuencia, se puede esperar que la discriminación entre los estados emocionales neutral y felicidad debe ser superior a la discriminación entre los estados triste y neutral.

Los resultados discutidos en esta sección sugieren que la similitud de contornos de F0 extraídos de la señal de test y las elocuciones de referencia puede ser utilizada para detectar el estado emocional en condiciones léxico dependiente y con señales generadas por un mismo hablante. Para eliminar estas restricciones, las secciones 4.4 y 4.5 proponen entrenar un *template* o plantilla de referencia neutro con elocuciones pronunciadas por varios locutores y con diferentes contenido léxico. Esta plantilla de referencia neutra puede ser implementada con *functional data analysis* (FDA).

## **4.4. Análisis de la prominencia emocional usando una familia de funciones**

### **4.4.1. Extensión del enfoque propuesto para modelar la variabilidad inter-locutor e intra-locutor en el contorno de F0**

En esta sección se construyen modelos de referencia neutros usando una familia de curvas de F0 mediante *functional PCA*. El esquema que se presenta aquí es inde-



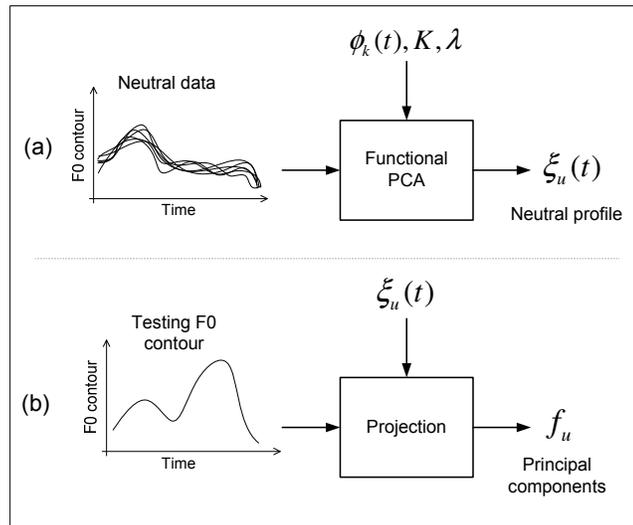
**Figura 4.1:** Distribución de la medida de similitud con las emociones neutra (neutral), enojo (angry), felicidad (happy) y tristeza (sad) en la base de datos EMA. La medida de similitud corresponde a la correlación de Pearson entre los contornos de F0 neutros y emocionales.

pendiente del locutor, pero aún léxico-dependiente. El análisis se realiza en el nivel de oración usando la base de datos EMA descrita en la sección 4.2.2.

La figura 4.2-a describe el marco general para construir la referencia neutra mediante el uso de *functional* PCA. En primer lugar, un conjunto de elocuciones neutras con el mismo contenido léxico pronunciadas por varios locutores se usan como datos de entrenamiento. Todas las señales están alineadas en el tiempo mediante DTW estándar. Luego, se aplica a las señales el procedimiento de extracción de F0 descrito en la sección 4.2.3. Las curvas de F0 alineadas y post-procesadas resultantes se suavizan y se representan como datos funcionales mediante el uso de una base de funciones  $\phi_k(t)$  *B-spline* de acuerdo a las ecuaciones (2.17) y (2.19). Finalmente, se aplica *functional*

PCA para generar una nueva base ortogonal de funciones  $\xi_u(t)$ .

La figura 4.2-b muestra la etapa de test del sistema propuesto. Como primer paso, la elocución de test se alinea con los datos de entrenamiento utilizando DTW. Luego se extrae el contorno de F0 y se estiman las proyecciones de curva de F0 de test sobre las base de funciones de referencia neutra  $\xi_u(t)$ . Como resultado, se obtienen los coeficientes  $f_u$ , que corresponden a parámetros que describen la forma del contorno de F0 de test. Dado que el perfil  $\xi_u(t)$  se genera con voz no emocional, se espera que los contornos de F0 neutros generarán proyecciones diferentes (es decir,  $\{f_1 \dots f_U\}$ ) sobre la base de funciones  $\xi_u(t)$ . Por lo tanto, el conjunto de parámetros  $\{f_1 \dots f_U\}$  podría ser utilizado para detectar emociones en la voz.

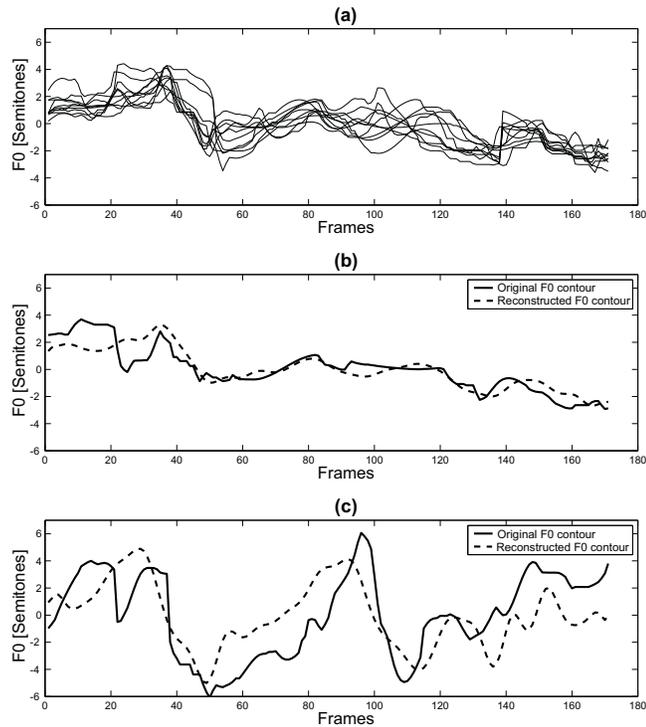


**Figura 4.2:** Marco general del método propuesto: (a) generación de modelos neutrales usando Functional PCA; y, (b) proyección de una señal de test en el espacio neutro.

La Figura 4.3 presenta un ejemplo del método propuesto para la oración “emph I am talking about the same picture you showed me” (extraída de la base de datos EMA).

La figura 4.3-a muestra las curvas de F0 alineadas en el tiempo y post-procesadas para diez realizaciones neutras pronunciadas por los locutores hablantes *ab* y *jn* (cinco repeticiones cada uno). A pesar de que las oraciones presentan variaciones en sus contornos de F0, es claro que tienen un patrón que el enfoque propuesto tiene como objetivo capturar. Este resultado coincide con trabajos previos que han demostrado que cuando el contenido léxico se mantiene constante, se obtienen mejoras en la exactitud de clasificación de emociones. Después, se entrena un perfil neutro con estos datos aplicando el procedimiento presentado en la figura 4.2-a. Para este ejemplo, la base suavizado  $\phi_k$  se implementa con una base sexta función de orden *B-spline* con  $K = 40$ . Las figuras 4.3-b y 4.3-c muestran la reconstrucción de del contorno de F0 de una elocución neutra y feliz, respectivamente, en la misma oración pronunciada por el locutor *ls* (no considerado para la construcción de la referencia neutra). Ambas curvas de F0 se reconstruyen utilizando las cinco primeras componentes principales. Como se puede apreciar en las Figs. 4.3-b y 4.3-c, el contorno de F0 neutro se aproxima con mayor exactitud que el contorno de F0 correspondiente a la oración emocionalmente feliz.

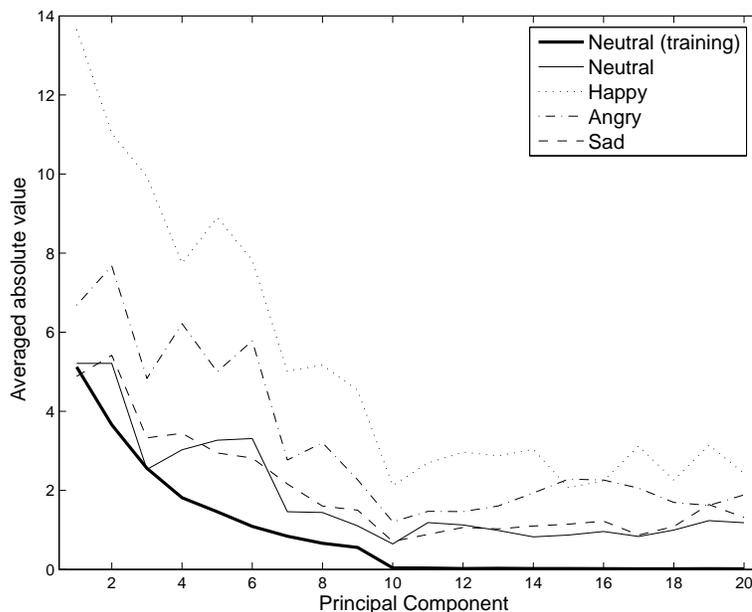
De acuerdo con las figuras 4.3-b y 4.3-c, el modelo de referencia se ajusta mejor a la frase neutra que el correspondiente a la oración feliz. Por lo tanto, es razonable concluir que la proyección sobre la  $k$ -ésima función base, con  $6 \leq k \leq 40$ , converge a cero más rápido con señales neutras que con elocuciones felices. En consecuencia, este análisis sugiere que las proyecciones de los contornos de F0 de la voz emocional son diferentes a aquellas generadas a partir de expresiones neutras. Este resultado es



**Figura 4.3:** Reconstrucción de los contornos de  $F_0$  usando *Functional PCA*: (a) datos de entrenamiento para generar la base neutral usando *functional PCA*; (b) reconstrucción de una señal de test neutra con las cinco primeras componentes principales; y, (c) reconstrucción de una elocución “feliz” (happy) usando las cinco primeras componentes principales. El error cuadrático medio entre el contorno de  $F_0$  original y reconstruido es igual a 0,45 y 0,32 para las señales neutra y feliz, respectivamente.

respaldado por la figura 4.4 que muestra el valor absoluto promedio de las proyecciones sobre los primeras 20 componentes principales para señales neutras y emocionales. Las proyecciones fueron generadas usando la estrategia *leave-one-out*. Los modelos de *functional PCA* fueron entrenados con dos locutores y testeados con el tercero. Como se puede ver en la figura 4.4, el promedio de las proyecciones de los contornos de  $F_0$  neutros es aproximadamente igual a cero cuando  $k \geq 10$ . El promedio del valor absoluto

de las proyecciones para la voz feliz y enojada es mayor que en el caso del habla neutra, incluso para las componentes principales de alto orden.



**Figura 4.4:** Valor absoluto promedio de las proyecciones asociadas a cada componente principal obtenida con la base de datos EMA.

#### 4.4.2. Análisis discriminante

Para evaluar el poder de discriminación de las proyecciones generadas con *functional PCA*, el sistema descrito en la figura. 4.2 es usado en una tarea de clasificación binaria entre señales neutras y emocionales. Este análisis tiene como objetivo validar la hipótesis de que las referencias neutras sobre basadas en FDA pueden ser utilizadas para detectar la prominencia emocional en la voz. Además, el método propuesto se compara con una técnica estándar de detección de emoción ampliamente usada en la literatura. Dado que no se necesitan múltiples repeticiones de las frases de cada locutor

como en la sección 4.3, el análisis considera tanto la base de datos EMA como el corpus EMO-DB.

La base de datos EMA se dividió en los siguientes conjuntos: desarrollo (la construcción de los modelos de referencia funcionales PCA), entrenamiento (para entrenar el clasificador) y test (para evaluar la exactitud del sistema). Cada uno de estos tres conjuntos contiene muestras de voz de un solo locutor, esto es, un locutor para estimar los modelos de referencia texto dependientes, un segundo hablante para entrenar el clasificador y un tercero para evaluar el sistema. Para construir los modelos de referencia, sólo se utilizaron datos neutrales. Para maximizar el uso de la base de datos de EMA, se realizaron seis permutaciones intercambiando el rol de cada locutor entre desarrollo, entrenamiento y test. Este procedimiento asegura que los resultados son independientes del locutor. La tasa de exactitud se calcula promediando los resultados obtenidos en las seis implementaciones. Se utilizó un clasificador QDC (*quadratic discriminant classifier*) para reconocer voz neutra y emocional (alegría, enojo y tristeza). El clasificador QDC calcula el *score* de salida mediante el uso de una combinación cuadrática del vector de características:  $y = x^T a x + b^T x + c$ , donde  $x$  e  $y$  son el vector de entrada y el *score* de salida, respectivamente. A pesar de que los clasificadores no lineales permiten obtener mejores resultados, se elige QDC por razones de simplicidad y generalización. Los clasificadores binarios neutro-felicidad, neutro-enojo, y neutro-tristeza fueron entrenados individualmente. Adicionalmente, se generó una cuarta clase agrupando las señales de las tres clases emocionales consideradas, denominada clase emocional. Cin-

cuenta elocuciones emocionales de la categoría emocional fueron escogidas al azar de modo tal que el número de muestras neutras coincida con el número de muestras emocionales (*chance* = 50 %). Este procedimiento se repitió 100 veces y se promediaron las tasas de rendimiento. Un procedimiento similar se llevó a cabo para la base de datos de EMO-DB, donde las señales se dividieron en subgrupos de desarrollo, entrenamiento y test.

La Tabla 4.2 muestra el rendimiento del sistema propuesto. Para la base de datos EMA, la exactitud en la clasificación neutro-emocional es igual a 91,3 %. Además, las exactitudes en la clasificación neutro-enojo y neutro-emocional son superiores al 75 %. Estos resultados validan el método propuesto. Para la base de datos DB-EMO, la exactitud de clasificación neutral-feliz y neutral-enojo están sobre el 73 %. Como era de esperar, la exactitud en la clasificación neutro-tristeza es baja para ambas bases de datos (EMA 63,3 %, EMO-DB 68 %). Estos resultados son consistentes con el análisis presentado en las secciones 4.3 y 4.4.1 (Figs. 4.1 y 4.4) donde se muestra que la discriminación entre las clases neutral y tristeza es menor que en los casos neutral-felicidad y neutral-enojo.

Con fines de comparación, se implementó un sistema de referencia en el estado del arte para detección de emociones binario que usa las estadísticas del F0 como características [77, 133, 137, 4]. En primer lugar, 80 funcionales a nivel de oración derivadas de F0 fueron extraídos usando la herramienta openSMILE [148]. El conjunto de funcionales corresponde a los mismos usados para el Interspeech 2010 paralinguistic challenge

**Tabla 4.2:** *Análisis discriminante para las proyecciones obtenidas con funcional PCA a nivel de oración usando bases léxico dependientes con las bases de datos EMA y EMO-DB (Acc = Accuracy, Pre = Precision, Rec = Recall, F = F-score). Chance corresponde al número total de muestras emocionales dividido por el número total de señales.*

		Acc	Pre	Rec	F	Chance
EMA	Neutral-Happy	0.913 (0.033)	0.884	0.960	0.918	0.500
	Neutral-Angry	0.777 (0.071)	0.784	0.780	0.777	0.500
	Neutral-Sad	0.633 (0.067)	0.626	0.693	0.654	0.500
	Neutral-Emotional	0.758 (0.057)	0.785	0.726	0.752	0.500
EMO-DB	Neutral-Fear	0.709 (0.039)	0.715	0.706	0.707	0.500
	Neutral-Disgust	0.710 (0.039)	0.715	0.707	0.707	0.500
	Neutral-Happiness	0.736 (0.025)	0.742	0.725	0.733	0.500
	Neutral-Boredom	0.639 (0.051)	0.661	0.562	0.604	0.500
	Neutral-Sadness	0.680 (0.035)	0.718	0.598	0.646	0.500
	Neutral-Anger	0.738 (0.013)	0.738	0.740	0.738	0.500
	Neutral-Emotional	0.699 (0.011)	0.715	0.663	0.687	0.500

[149]. Luego, se aplicó *forward feature selection* (FFS) para reducir el número de funcionales a 20, igualando el número de proyecciones utilizadas como características en el método propuesto en este trabajo. Un clasificador QDC también fue implementado con el sistema de referencia. Para la base de datos de EMA, los clasificadores fueron entrenados con dos locutores y testeado con un tercer hablante. Tres permutaciones fueron generadas intercambiando los roles de cada locutor. Las mismas cuatro clases emocionales que en el caso del sistema propuesto fueron definidas (alegría, enojo, tristeza y emocional). Este experimento siguió el mismo procedimiento adoptado para los resultados que se muestran en la Tabla 4.2. Del mismo modo, un sistema de referencia se construyó para la base de datos de EMO-DB. La Tabla 4.3 muestra los resultados de los experimentos para este sistema de referencia en el estado del arte, tanto para el corpus EMA como la base de datos EMO-DB.

En la base de datos de EMA, la exactitud del sistema propuesto en la clasifica-

**Tabla 4.3:** Desempeño del sistema de referencia con las bases de datos EMA y EMO-DB. Las características fueron extraídas a partir del contorno de F0 a nivel de oración (Acc = Accuracy, Pre = Precision, Rec = Recall, F = F-score). Chance corresponde al número total de muestras emocionales dividido por el número total de señales.

		Acc	Pre	Rec	F	Chance
EMA	Neutral-Happy	0.843 (0.086)	0.927	0.780	0.819	0.500
	Neutral-Angry	0.737 (0.120)	0.882	0.627	0.666	0.500
	Neutral-Sad	0.653 (0.195)	0.683	0.753	0.687	0.500
	Neutral-Emotional	0.714 (0.116)	0.723	0.762	0.732	0.500
EMO-DB	Neutral-Fear	0.642 (0.121)	0.861	0.370	0.469	0.500
	Neutral-Disgust	0.658 (0.108)	0.726	0.562	0.583	0.500
	Neutral-Happiness	0.763 (0.094)	0.964	0.552	0.679	0.500
	Neutral-Boredom	0.513 (0.016)	0.875	0.055	0.102	0.500
	Neutral-Sadness	0.706 (0.087)	0.644	0.966	0.769	0.500
	Neutral-Anger	0.825 (0.120)	0.930	0.709	0.774	0.500
	Neutral-Emotional	0.690 (0.097)	0.889	0.458	0.555	0.500

ción neutro-felicidad es igual a 7.0 % (absoluto) más alto que el sistema de referencia (estadísticamente significativa con  $p$ -valor= 0,001 , ver Tabla 4.4). Además, los clasificadores neutro-enojo y neutro-emocional logran mejoras de 4.0 % y el 4.4 % (absoluto), respectivamente, en comparación con el método de referencia ( $p$ -valor= 0,092 y  $p$ -valor = 0,078, respectivamente, ver Tabla 4.4). En la base de datos EMO-DB, el método propuesto conduce a un aumento en la exactitud en la clasificación neutro-miedo, neutro-asco y neutro-aburrimiento igual al 6.8 %, 5.2 % y 12,6 % (absoluto), respectivamente. Estos resultados sugieren que el esquema propuesto puede discriminar con precisión entre las categorías neutral y emocional. Sin embargo, la precisión alcanzada por el sistema presentado en este trabajo es más baja para la clasificación neutro-tristeza en las bases de datos EMA (2.0 %) y EMO-DB (2.6 %). Cabe destacar que, en comparación con el método de referencia, las mejoras en exactitud alcanzadas por el sistema propuesto en clasificación neutro-emocional, neutro-felicidad y neutro-rabia son mucho

más altas que la degradación en exactitud en la clasificación neutro-tristeza. Por otra parte, las desviaciones estándar de la exactitud dadas por el sistema propuesto (Tabla 4.2) son mucho más bajas que aquellas obtenidas con el método de referencia (Tabla 4.3). Estos resultados sugieren que el clasificador basado en *funcional* PCA es más confiable y consistente que el método de referencia.

**Tabla 4.4:** *Test de hipótesis (proporciones) para determinar si las diferencias entre los clasificadores son estadísticamente significativas en la base de datos EMA. Los colores claros y oscuros representan significancia estadística fuerte ( $p\text{-value} < 0.05$ ) y débil ( $p\text{-value} < 0.1$ ), respectivamente (R = Sistema de Referencia, LD = modelos léxico-dependientes, LI = modelos léxico-independientes, Vf = segmentación basada en ventanas de tamaño fijo, Ch = Segmentación a nivel de chunk, Pa = segmentación a nivel de palabras)*

	Happiness	Anger	Sadness	Emotion
LD - R	0.001	0.092	0.278	0.078
LI - R	0.001	0.088	0.283	0.081
LD - LI	0.121	0.219	0.001	0.266
LI - Vf	0.228	0.431	0.425	0.304
LI - Ch	0.041	0.431	0.030	0.001
LI - Pa	0.001	0.018	0.015	0.190

### 4.4.3. Bases léxico-dependiente versus léxico-independiente

Los resultados presentados en la sección 4.4.2 muestran que el enfoque basado en FDA propuesto en este trabajo puede discriminar con precisión entre el voz neutra y emocional. Sin embargo, un sistema de detección de emociones léxico dependiente no es factible de utilizar en aplicaciones reales. En esta sección se generaliza el método propuesto para el caso léxico independiente. Básicamente, la idea es construir la base de funciones utilizando frases neutras que transmiten información léxica diferente. Los resultados de la sección 4.4.2 mostraron que la información léxica afecta el contorno

de F0, incluso para las lenguas no tonales. Las bases independientes del léxico no capturarán este aspecto. Sin embargo, al relajar la restricción léxico-dependiente, se pueden utilizar más frases para construir la base de funciones para *functional* PCA. De esta forma, será posible construir modelos de referencia robustos que capturen de mejor manera la variabilidad de F0.

En primer lugar se extrae el F0 de las elocuciones neutrales léxico independientes y se post-procesan de acuerdo al método descrito en la sección 4.2.3. Después, se calcula la duración media de las señales la cual se utiliza para deformar linealmente en el tiempo los contornos de F0. La familia de curvas resultante se utiliza como entrada al sistema basado en *functional* PCA (Figura 4.2-a). Para evaluar el desempeño de la técnica léxico-independiente, se realiza un análisis discriminante de la misma manera como se describe en la sección 4.4.2. Observar que en la sección 4.4.1 había 10 modelos neutros basados en *functional* PCA dependientes del texto (es decir, uno por cada oración). Ahora, sólo hay un modelo independiente del léxico el cual fue entrenado con todos los contornos de F0 neutros.

La Tabla 4.5 muestra los resultados de clasificación usando proyecciones basadas en *functional* PCA con modelos léxico-independientes para las bases de datos EMA y EMO-DB. En el caso del corpus EMA, la exactitud alcanzada en clasificación neutro-felicidad y neutro-emocional son sólo 2,0% (absoluto) y 1,6% (absoluto) más bajos que la exactitud obtenida usando modelos léxico-dependientes (véase las Tablas 4.2 y 4.5). Para la base de datos EMO-DB, la exactitud en la clasificación neutro-emocional

**Tabla 4.5:** *Análisis discriminante para las proyecciones obtenidas con funcional PCA a nivel de oración para bases léxico-independientes con las bases de datos EMA y EMO-DB (Acc = Accuracy, Pre = Precision, Rec = Recall, F = F-score). Chance corresponde al número total de muestras emocionales dividido por el número total de señales.*

		Acc	Pre	Rec	F	Chance
EMA	Neutral-Happy	0.893 (0.059)	0.874	0.933	0.899	0.500
	Neutral-Angry	0.795 (0.059)	0.793	0.827	0.802	0.500
	Neutral-Sad	0.547 (0.063)	0.550	0.573	0.553	0.500
	Neutral-Emotional	0.742 (0.057)	0.784	0.701	0.733	0.500
EMO-DB	Neutral-Fear	0.709 (0.050)	0.737	0.673	0.685	0.500
	Neutral-Disgust	0.711 (0.049)	0.738	0.675	0.686	0.500
	Neutral-Happiness	0.789 (0.032)	0.788	0.806	0.793	0.500
	Neutral-Boredom	0.706 (0.057)	0.753	0.614	0.674	0.500
	Neutral-Sadness	0.663 (0.056)	0.807	0.432	0.557	0.500
	Neutral-Anger	0.777 (0.034)	0.780	0.785	0.779	0.500
	Neutral-Emotional	0.713 (0.036)	0.756	0.641	0.691	0.500

que logra el sistema con modelos léxico independientes es 1,5 % (absoluto) mayor que la exactitud alcanzada con los modelos de léxico dependientes. De acuerdo con un test de hipótesis para proporciones, estas diferencias no son estadísticamente significativas (ver la Tabla 4.4). Además, cuando se compara con el sistema de referencia (Tabla 4.3), el sistema léxico-independiente basado en PCA conduce a mejoras en la exactitud iguales a 5,0 % (absoluto), 5,8 % (absoluto) y 2,8 % (absoluto) para la clasificación neutro-felicidad, neutro-enojo y neutro-emocional, respectivamente (base de datos EMA). Todas estas diferencias son estadísticamente significativas (véase Tabla 4.4). Resultados similares se obtienen con la base de datos de EMO-DB. En efecto, el sistema léxico independiente propuesto conduce a mejoras en la exactitud iguales a 6,7 %, 5,2 % y el 2,3 % con neutro-miedo, neutro-asco y neutro-emocional.

## 4.5. Análisis y evaluación de la prominencia emocional a nivel de sub-oración

En esta sección se generaliza la técnica léxico-independiente basado en *functional* PCA (véase la sección 4.4.3) a nivel de sub-frase (por ejemplo, *chunk* o palabra). La prominencia emocional transmitida en el contorno de F0 no se distribuye uniformemente en el tiempo [77, 96, 76]. Al extender el análisis a sub-unidades de la oración, el objetivo es detectar aquellos segmentos emocionalmente importantes. Este enfoque no requiere de dividir un diálogo en oraciones. Por lo tanto, es aplicable para los sistemas de detección de emociones en tiempo real.

El sistema basado en *functional* PCA de detección de emociones que se muestra en la Fig. 4.2 se aplica a tres sub-unidades diferentes: ventanas de duración definida; frase (o *chunk*); y, palabra. La segmentación basada en ventanas de duración fija consiste en dividir la señal de voz en las ventanas de un segundo con el 50% de traslape [150]. Esta segmentación no requiere estimar los límites sintácticos de las elocuciones. Una frase o *chunk* se define como un grupo de palabras que conforman una única unidad sintaxis, que a su vez es adecuada para el reconocimiento de emociones [146]. Los avances recientes en procesamiento del lenguaje han proporcionado herramientas para dividir automáticamente una oración dada en frases. En este trabajo se utiliza un identificador de *chunks* basado en SVM (*support vector machines*) propuesto por Kudoh y Matsumoto [151]. La segmentación a nivel de palabra también se incluyó en el análisis, a pesar de que su longitud puede no ser suficiente para capturar la información

suprasegmental. La segmentación de palabras se obtiene mediante alineamiento forzado basado en modelos oculto de Markov (HMM, *hidden markov models*) [143].

Se utiliza un corpus neutro para construir los modelos neutros léxico-independientes a nivel de sub-oración (ventana de duración fija, *chunk* o palabra). Este corpus es *Wall Street Journal Continuous Speech Recognition Corpus Phase II* (WSJ1) [152] (véase la Tabla 4.1). Esta base de datos considera 8104 elocuciones espontáneas grabadas por 50 locutores con diversos grados de experiencia en dictado. En primer lugar, las señales fueron segmentados de acuerdo a cada tipo de sub-frase. Después, 200 señales fueron escogidas al azar entre el total de 50 locutores para extraer los segmentos correspondientes (ventanas de duración fija, *chunk* o palabra). Estas 200 elocuciones dan origen a más de 1500 segmentos neutrales para cada uno de los niveles de segmentación, los que fueron utilizados para construir las bases usando *functional* PCA. Para cada tipo de unidad sub-oración, los contornos de F0 son linealmente deformados de modo tal que su duración alcance el promedio de todos los segmentos. El conjunto resultante de contornos de F0 se utiliza como entrada al sistema basado en *functional* PCA descrito en la Fig. 4.2-a. Esta evaluación se realizó utilizando las bases de datos de EMA y EMO-DB. Dado que no está disponible la segmentación a nivel de palabras para la base de datos de EMO-DB, sólo se informan los resultados para segmentación basada en ventanas de duración fija.

Las elocuciones de test se segmentan de acuerdo a las unidades sub-oración descritas anteriormente. Después de la extracción y el post-procesamiento de sus contornos de

F0, se calculan las proyecciones sobre la base obtenida con *functional* PCA (Fig. 4.2-b) para cada tipo de segmento. Los clasificadores QDC, que se entrenan para cada nivel de sub-oración, se utilizan para clasificar cada segmento en la señal de test utilizando un esquema *leave-one-out*. Por último, la clase emocional a a nivel de oración se calcula promediando los *scores* del clasificador QDC dentro de la elocución de test [153].

**Tabla 4.6:** *Exactitud (accuracy) para diferentes niveles de segmentación usando las bases de datos EMA y EMO-DB con modelos léxico-independientes. En el corpus EMO-DB, los resultados para los niveles Chunk y palabra no se entregan dado que la segmentación a nivel de fonemas no está disponible.*

		Nivel de segmentación			
		Oración	Ventana fija	Chunk	Palabra
EMA	Neutral-Happiness	0.893	0.877	0.853	0.817
	Neutral-Anger	0.795	0.790	0.790	0.733
	Neutral-Sadness	0.547	0.553	0.480	0.470
	Neutral-Emotional	0.742	0.744	0.773	0.745
EMO-DB	Neutral-Fear	0.709	0.580	-	-
	Neutral-Disgust	0.711	0.750	-	-
	Neutral-Happiness	0.789	0.694	-	-
	Neutral-Boredom	0.706	0.595	-	-
	Neutral-Sadness	0.663	0.722	-	-
	Neutral-Anger	0.777	0.706	-	-
	Neutral-Emotional	0.713	0.744	-	-

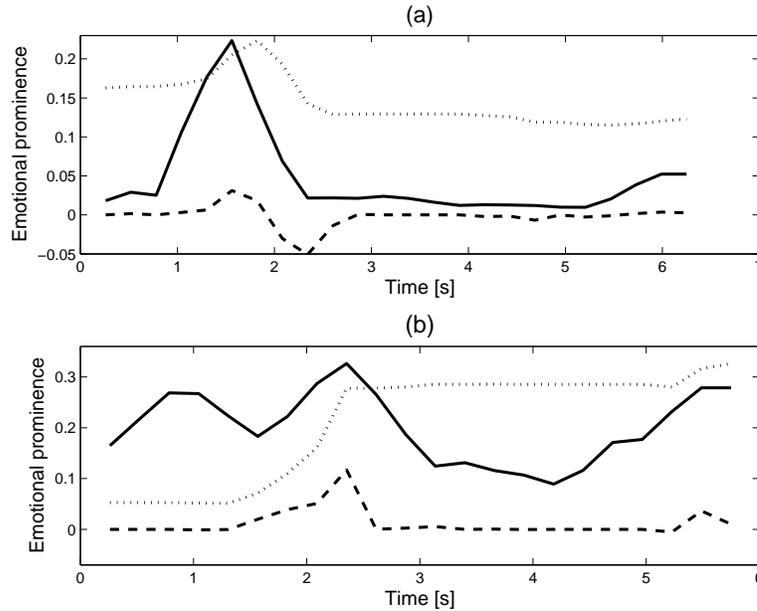
**Tabla 4.7:** *Inter-evaluator agreement (IEA) y correlaciones subjetiva-objetiva con la base de datos SEMAINE para diferentes tamaños de ventana (IEA = correlación entre la evaluación subjetiva de un sujeto y el promedio de los restantes evaluadores en la base de datos,  $\rho(S, O)$  = correlación entre el promedio de las evaluaciones subjetivas y la métrica objetiva,  $\rho(\frac{\Delta S}{\Delta t}, O)$  = correlación entre la derivada del promedio de las evaluaciones subjetivas y la métrica objetiva).*

Window length [s]	IEA	$\rho(S, O)$	$\rho(\frac{\Delta S}{\Delta t}, O)$
0.25	0.347	0.235	0.327
0.50	0.351	0.250	0.396
1.00	0.363	0.215	0.438

La Tabla 4.6 presenta la exactitud promedio para la segmentación basada en ven-

tanías de duración fija, *chunk* y segmentación a nivel de palabra. También se muestran los resultados a nivel de la oración con modelos de referencia léxico independientes (valores extraídos de la Tabla 4.5). La Tabla 4.6 muestra que, en general, la exactitud lograda a nivel de oración es superior a la alcanzada a nivel de sub-oración. Entre los diferentes tipos de segmentación, aquella basada en ventanas de duración fija proporciona la más alta exactitud de clasificación emocional. Como se puede observar en la Tabla 4.4, las diferencias en la exactitud de clasificación entre el nivel de oración y la segmentación basada en ventanas no es significativa para todas las categorías (base de datos EMA). Se observan resultados similares con la base de datos de EMO-DB. Sin embargo, las diferencias más significativas en la exactitud se observan cuando se comparan los clasificadores a nivel de oración y a nivel de palabras. Este resultado es consistente trabajos anteriores, los cuales sugieren que las unidades cortas en duración no son eficaces para capturar la información emocional a partir del contorno de F0 [77].

El enfoque propuesto es validado con la base de datos espontánea del proyecto SEMAINE [135] (véase la Tabla 4.1 y la sección 4.2.2). En lugar de asignar una etiqueta emocional a cada oración, las evaluaciones subjetivas corresponden a la evaluación continua del contenido emocional en tiempo real utilizando la herramienta Feeltrace (diez valores por segundo). Por lo tanto, esta base de datos es ideal para evaluar si el método propuesto puede detectar en forma localizada la información emocional transmitida en la voz. En la literatura hay varios autores que han considerado esta base de datos para reconocer valores altos y bajos de *valence*, *arousal*, *expectancy* y *power*



**Figura 4.5:** (a) Ejemplo de la métrica subjetiva (punteada), derivada de la subjetiva (segmentada) y objetiva (continua). En este ejemplo, la correlación entre las métricas objetiva y subjetiva es igual a  $\rho = 0,51$ . (b) Ejemplo de la métrica subjetiva (punteada), derivada de la subjetiva (segmentada) y objetiva (continua). En este ejemplo, la correlación entre las métricas objetiva y la derivada de la métrica subjetiva es igual a  $\rho = 0,55$ .

[154, 155, 156]. Ellos han reportado exactitudes en torno al 50% en clasificación binaria a nivel de palabras. Como las configuraciones usadas por estos autores son diferentes a aquellas mostradas en este trabajo, los resultados no se pueden comparar directamente. Sin embargo, estos estudios muestran lo complejo que es reconocer el estado emocional en este corpus.

Esta evaluación se compara la similitud entre una métrica objetiva derivada de las proyecciones basadas en *funcional* PCA y los valores promedio de las evaluaciones subjetivas. Los experimentos se realizaron usando segmentación basada en ventanas de duración fija (explicada anteriormente). Se utilizó la base de datos WSJ1 para

construir la base neutral léxico independiente. Para cada ventana, se calcularon las proyecciones en esta base. En virtud de los resultados de la figura 4.4, se calculó la norma de la proyección, que se utiliza como una métrica objetiva de la prominencia emocional transmitida en la voz. Esta norma se suaviza con un filtro de mediano. Como *ground truth* se estimó una medida subjetiva de la prominencia emocional. Mientras se utiliza Feeltrace, los evaluadores tienen instrucciones de poner el cursor en el centro del sistema de coordenadas para describir el estado neutro. La distancia del puntero desde el centro se considera que la intensidad emocional de la voz [97]. Por lo tanto, definimos la métrica subjetiva,  $e(t)$ , como:

$$e(t) = \sqrt{a^2(t) + v^2(t)} \quad (4.2)$$

donde  $a(t)$  y  $v(t)$  son las curvas de *activation* y *valence* promedio, respectivamente, dada por los evaluadores humanos.

El uso de Feeltrace como herramienta de evaluación de la prominencia emocional presenta algunos desafíos. Los evaluadores tienen que percibir los estímulos, percibir el mensaje, deben identificar sus atributos emocionales y mover el puntero de acuerdo a su juicio de percepción, todo esto en tiempo real. El proceso de percepción introduce un retardo que es intrínsecamente independiente del hablante [157]. Sin embargo, el enfoque propuesto captura el contenido emocional de la señal de manera instantánea. El desfase entre las señales se aborda usando un retardo entre las evaluaciones objetivos y subjetivas. Este retardo se calcula para cada elocución mediante la maximización

de la correlación entre las métricas. El retardo es forzado a ser menor a 0,5 segundos (Nicolaou *et al.* proponen un umbral similar [158] para abordar el problema de sincronización).

La Tabla 4.7 presenta el promedio de correlación de Pearson entre las medidas objetivas y subjetivas que describen la prominencia emocional en la base de datos SEMAINE (columna  $\rho(S, O)$ ). Estos resultados se estiman usando segmentación basada en ventanas de duración fija (0,25, 0,5 y 1 seg), con un 50 % de traslape. La correlación promedio entre las medidas objetivas y subjetivas es de  $\rho = 0,25$  cuando el tamaño de la ventana es igual a 0,5 seg. A modo de comparación, la Tabla 4.7 también muestra el *inter-evaluator agreement* (IEA), que corresponde a la correlación promedio entre las curvas de un evaluador y las curvas promedio de los otros evaluadores (columna *IEA*). El IEA es de  $\rho = 0,35$  cuando el tamaño de la ventana se ajusta en 0,5 segundos. A pesar de que el método presentado en este trabajo ofrece una menor correlación, la métrica propuesta se acerca a la correlación observada entre los evaluadores. Se debe tener en cuenta que esta comparación no es del todo justa, ya que los evaluadores realizaron las evaluaciones después de ver los vídeos y escuchar el audio. En cambio la métrica objetiva propuesta se estimó usando únicamente los contornos de F0. La Figura 4.5-a muestra un ejemplo de métricas objetivas y subjetivas para una elocución dada en la base de datos SEMAINE (el eje  $x$  corresponde al tiempo). Por razones de visualización, se aplicó un factor de normalización constante a la medida objetiva. Para este ejemplo, la correlación entre ambas curvas es  $\rho = 0,51$ . En esta figura también se

puede apreciar el desfase entre ambas mediciones.

En algunas elocuciones se observó que la correlación entre las mediciones objetivas y subjetivas es baja o incluso negativa. La Figura 4.5-b muestra un ejemplo en el que la correlación entre las curvas es  $\rho = -0,24$ . Un patrón interesante en esta figura es el comportamiento acumulado de la curva subjetiva, que también se observó en otras señales. La hipótesis propuesta es que, después de percibir un segmento localizado de alta intensidad emocional, los evaluadores humanos tienden a mantener la posición del cursor en el mismo lugar desde por algún tiempo a pesar de que la intensidad emocional intrínseca disminuye. Este comportamiento acumulado también ha sido observado por otros autores quienes muestran que los individuos son más sensibles a las variaciones relativas en la intensidad emocional [159]. De hecho, la mayor variación en la curva subjetiva se muestra en la Figura 4.5-b coincide con la máxima prominencia emocional registrada por la medición objetivo propuesta. Tendiendo en cuenta estos resultados, se ha tomado la determinación de comparar la métrica objetiva propuesta con la derivada de las curvas subjetivas (es decir, las variaciones en lugar de los valores absolutos). Por ejemplo, la Figura 4.5-b muestra la derivada de la evaluación subjetiva (línea segmentada). La correlación entre esta señal y la métrica propuesta es de  $\rho = 0,55$ . La Tabla 4.7 muestra la correlación entre la métrica objetiva propuesta y la derivada de la medida subjetiva para todas las señales (columna  $\rho(\frac{\Delta S}{\Delta t}, S)$ ). Las correlaciones son más altas que cuando se utilizan los valores absolutos de las curvas subjetivas (columna  $\rho(S, O)$ ). Curiosamente, los valores de correlación son aún mayores que el

*inter-evaluator agreement*, cuando el largo de la ventana es igual a 0,5 o 1 segundo.

## 4.6. Conclusiones

Este capítulo propone un método para detectar la modulación emocional en los contornos de F0 mediante el uso de modelos de referencia, definidos neutrales, basados en *functional* PCA. La técnica propuesta también puede ser usada para detectar los segmentos emocionalmente más sobresalientes dentro de una elocución. En primer lugar, se evalúa el esquema basado en la comparación de una señal emocional y una sola elocución neutral de referencia, ambas con el mismo contenido léxico y pronunciada por el mismo locutor. Los experimentos con las condiciones léxico-dependiente y locutor-dependiente sugieren que es factible emplear referencias neutrales para detectar modulación emocional en contornos de F0. En segundo lugar, la condición locutor-dependiente se eliminó mediante el uso de una base de funciones entrenada con contornos de F0 provenientes de más de un locutor usando *functional* PCA. El sistema propuesto con las condiciones locutor-independiente pero aún léxico-dependiente logra una exactitud tan alta como 75,8% en clasificación binaria, que a su vez es un 6,2% más alta que la obtenida con un detector de emociones estándar basado en estadísticas de F0. En tercer lugar, la condición léxico-dependiente es eliminada y la base de funciones se entrenó con contornos de F0 extraídos de las señales con diferente contenido léxico y pronunciadas por más de un locutor. Los resultados muestran que la degradación de la exactitud proporcionada por la técnica propuesta con modelos léxico-independientes

no es significativa cuando se compara con el sistema léxico-dependiente (esto es, de 75,8 % a 74,2 %). Por último, el sistema propuesto se aplica a nivel de sub-oración para detectar los segmentos emocionalmente más importantes. Cuando el método presentado en este capítulo se utiliza a nivel de sub-oración, la diferencia en exactitud a nivel de oración y a nivel de segmentos usando ventana de duración fija no es significativa para todas las categorías emocionales. Además, los experimentos con una base de datos espontánea muestra que la correlación entre la derivada de las evaluaciones subjetivas y las objetivas entregadas por el sistema propuesto es igual a  $\rho = 0,44$  (*inter-evaluator agreement*  $\rho = 0,36$ ).

El trabajo futuro incluye la incorporación de otras características prosódicas (esto es, la energía y la duración) y parámetros espectrales (por ejemplo, MFCCs) en base de funciones basada en *functional PCA*. Además, el enfoque actual puede ser extendido al análisis de descriptores faciales para detección de emociones. Del mismo modo, el método puede ser utilizado para detectar categorías emocionales específicas (por ejemplo, alegría versus enojo). Por ejemplo, es posible construir bases de funciones para una emoción específica. Finalmente, el método presentado puede extenderse incluso a otras tareas de procesamiento de voz como por ejemplo la evaluación de prosodia en enseñanza de idiomas (problema abordado en el capítulo 3). Básicamente, la idea es la construir una base de funciones usando muestras de voz generadas por un hablante nativo. Luego, los contornos de F0 de test pronunciados por locutores no nativos pueden ser evaluados usando las proyecciones en la base de funciones nativa.

# Capítulo 5

## Conclusiones

En este trabajo se abordaron dos problemas que involucran la modelación de la prosodia en señales de voz. Primero, se presenta una propuesta para evaluar en forma automática la entonación en enseñanza de idiomas basado en un esquema *top-down*. La técnica calcula una medida de similitud que resulta de la comparación entre una señal de referencia con la señal de test generada por el usuario. Los resultados obtenidos muestran una alta correlación entre las evaluaciones objetivas dadas por los sistemas y los puntajes dados por expertos. Además, las técnicas son robustas a *mismatch* de pronunciación a nivel de segmentos, lo que permite separar efectivamente la evaluación prosódica y la pronunciación a nivel de sonidos individuales. Además, dado que los métodos son independientes del texto y del idioma, pueden ser fácilmente incorporados en software educativo ad hoc facilitando su masificación. Por lo tanto, estos resultados sugieren que los sistemas propuestos pueden ser utilizados en sistemas reales. Vale la

pena destacar el impacto social y tecnológico de la propuesta de esta tesis en enseñanza de idiomas, ya que corresponde a un caso de investigación aplicada a un problema de interés general.

Como resultado de las técnicas para CALL se propuso un esquema novedoso basado en *functional data analysis* para detección de emociones en señales acústicas. El método utiliza el concepto de modelo de referencia neutral con el que se comparan las elocuciones de test para evaluar su estado afectivo. El modelo de referencia consiste en una base de funciones independiente del locutor e independiente del texto. La técnica propuesta fue evaluada para bases de datos actuadas, donde se muestran mejoras respecto a sistemas en el estado del arte basados en estadísticas globales. Estos resultados validan el marco basado en FDA para analizar voz afectiva desde el punto de vista de la prosodia. Luego se propuso un esquema a nivel de sub-oración que entregó altas tasas de correlación entre las métricas objetivas dadas por el sistema y los indicadores emocionales entregados por humanos en una base de datos real. Los resultados sugieren que el sistema puede ser utilizado en aplicaciones reales donde se requiera interfaces hombre-máquina sensibles a emociones.

En el caso de CALL, como trabajo futuro se propone la integración de los sistemas propuestos en este documento con técnicas de evaluación de la calidad de pronunciación a nivel de segmentos. Asimismo, se propone el uso de otras características suprasegmentales como la duración en el caso de evaluación de acento. En el caso de detección de emociones también se contempla incorporar al esquema propuesto otras características

prosódicas como la energía y la duración, así como también parámetros espectrales. Además, se propone usar el método presentado en este trabajo para detectar estados emocionales específicos. También se propone aplicar las técnicas presentadas al reconocimiento de emociones usando descriptores faciales. Finalmente, el método para detección de emociones también puede ser aplicado en evaluación de entonación. La idea es construir una base de funciones con suficientes muestras de voz provenientes de hablantes nativos y utilizar este modelo para determinar si la entonación de un locutor no nativo es correcta o no.

# Glosario

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

**Activación:** Característica asociada al dinamismo de una emoción (i.e. si es activa o pasiva).

**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 segmentales:** Sonidos individuales del habla que se relacionan con el lugar y la forma de articulación

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

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

**Conjunto de entrenamiento:** Grupo de señales utilizadas para determinar los parámetros que describen un modelo.

**Conjunto de test:** Grupo de señales usadas para evaluar un sistema de clasificación determinado. Este set de señales es distinto al conjunto de entrenamiento.

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

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

**EER:** *Equal Error Rate*.

**Emoción:** Estado natural e instintivo de la mente humana que se deriva de las circunstancias, humor y la relación con otros.

**Entonación:** Combinación de estructuras tonales relacionadas con la variación del parámetro F0 que puede reflejar diferencias de sentido, de intención, de emoción y de origen del locutor.

**FA:** Falsa aceptación.

**FDA:** *Functional Data Analysis*, conjunto de técnicas estadísticas para representar y analizar señales en el dominio de las funciones continuas.

**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 eventanado. Unidad mínima de análisis.

**Frecuencia fundamental F0:** 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.

**Pitch:** Percepción de la frecuencia fundamental F0.

**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:** Receiver Operating Characteristic.

**Sonido sonoro:** Sonido que involucra vibración de las cuerdas vocales.

**Sonido sordo o áfono:** Aquel sonido que no se genera a partir de la excitación de las cuerdas vocales.

**Tono:** Rasgo prosódico presente en las lenguas tonales relacionado que tiene funciones léxico-semánticas.

**Valencia:** Característica de una emoción que permite identificarla como positiva o negativa.

# Capítulo 6

## Referencias

# Referencias

- [1] E. Shriberg and A. Stolcke, “Prosody modeling for automatic speech recognition and understanding,” *Mathematical Foundations of Speech and Language Processing*, Springer-Verlag, New York, vol. 138, pp. 105–114, 2004.
- [2] M. Kockmann, L. Ferrer, L. Burget, E. Shriberg, and J. Cernocky, “Recent progress in prosodic speaker verification,” in *ICASSP 2011*, Prague, Czech Republic, May 1998, pp. 4556–4559.
- [3] R. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.
- [4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, “Emotion recognition in human-computer interaction,” *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, January 2001.
- [5] D. Zhang, *Automated Biometrics - Technologies and Systems*. Kluwer Academics Publishers, 2000.
- [6] A. Fox, *Prosodic Features and Prosodic Structure: The Phonology of Suprasegmentals*. Oxford University Press, 2000.

- [7] L. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb 1989.
- [8] J. Laver, *Principles of Phonetics*. Cambridge University Press, 1994.
- [9] A. Botinis, B. Granström, and B. Möbius, “Developments and paradigms in intonation research,” *Speech Communication*, vol. 33, no. 4, pp. 263–296, March 2001.
- [10] D. Chun, *Discourse Intonation in L2*. John Benjamins., 2002.
- [11] J. Pierrehumbert and J. Hirschberg, *The meaning intonational contours in English*. In Eds. P. Cohen, J. Morgan & M. Pollack, *Intentions in communication*, MIT Press, 1990.
- [12] P. Roach, *English Phonetics and Phonology*. 3rd Ed. Cambridge: Cambridge University Press, 2008.
- [13] A. Cruttenden, *Gimson’s Pronunciation of English*. 7th edn., London: Hodder Education, 2008.
- [14] J. Wells, *English Intonation*. Cambridge: Cambridge University Press, 2006.
- [15] D. Jones, *An Outline of English Phonetics*. Cambridge: W. Heffer & Sons Ltd., 1962.

- [16] L. Brosnahan and B. Malmberg, *Introduction to Phonetics*. Cambridge: W. Heffer & Sons Ltd., 1970.
- [17] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York, Harper & Row, 1968.
- [18] H. Vivanco, “Análisis fonético acústico de una pronunciación de “ch” en jóvenes del estrato medio-alto y alto de Santiago de Chile,” in *Boletín de Filología de la Universidad de Chile*, 1999.
- [19] J. Picone, “Signal modeling techniques in speech recognition,” in *Proceedings of the IEEE*, 1993, pp. 1215–1247.
- [20] P. Bagshaw, *Automatic prosodic analysis for computer aided pronunciation teaching*. PhD Thesis, The University of Edinburgh, 1994.
- [21] L. Rabiner, “On the use of autocorrelation analysis for pitch detection,” *IEEE Transactions Acoustic Speech, Signal Processing*, vol. 25, no. 1, pp. 613–625, 1977.
- [22] T. Shimamura and H. Kobayashi, “Weighted autocorrelation for pitch extraction of noisy speech,” *IEEE Transactions Acoustic Speech, Signal Processing*, vol. 9, no. 1, pp. 727–730, 2001.
- [23] W. Hung, “Use of fuzzy weighted autocorrelation function for pitch extraction from noisy speech,” *Electronic Letters*, vol. 38, no. 19, pp. 1148–1150, 2002.

- [24] A. de Cheveigné and H. Kawahara, “Yin, a fundamental frequency estimator for speech and music,” *Journal Acoustic Society Am., IRCAM.*, 2002.
- [25] R. MacAulay, “Maximum likelihood pitch estimation using state-variable techniques,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '78)*, April 1978, pp. 1215–1247.
- [26] Y. Cho, H. Kim, M. Kim, and S. Kim, “Joint estimation of pitch, band magnitudes, and v/uv decisions for mbe vocoder,” in *EUROSPEECH 1997*, 1997, pp. 1271–1274.
- [27] E. Barnard, R. Cole, M. Veà, and F. Alleva, “Pitch detection with a neural-net classifier,” *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 298–307, 1991.
- [28] A. Saffiotti and A. Soquet, “Pitch determination of speech signals: A fuzzy fusion approach,” in *7th IFSA World Congress*, Prague, Czech Republic, 1997, pp. 315–320.
- [29] H. Huang and J. Pan, “Speech pitch determination based on hilbert-huang transform,” *Signal Processing (Elsevier)*, vol. 86, no. 4, pp. 792–803, 2006.
- [30] I. Gavati, M. Zirra, and O. Cula, “Intonation estimation for romanian language,” in *INT-1997*, Athens, Greece, September 1997, pp. 149–152.

- [31] J. Droppo and A. Acero, “Maximum a posteriori pitch tracking,” in *ICSLP-1998*, Sydney, Australia, November 1998.
- [32] C. Wang and S. Seneff, “Robust pitch tracking for prosodic modeling in telephone speech,” in *ICASSP 2000*, Istanbul, Turkey, August 2000, pp. 1343–1346.
- [33] K. Kasi and S. Zahorian, “Yet another algorithm for pitch tracking,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’02)*, Orlando, FL, USA, May 2002, pp. 361–364.
- [34] L. Rabiner, M. Sambur, and C. Schmidt, “Application of a nonlinear smoothing algorithm to speech processing,” *IEEE Transactions on Speech and Audio Processing*, vol. ASSP-23, 1975.
- [35] K. Yong-Duk-Cho, K. Al-Naimi, and A. Kondoz, “Pitch post-processing technique based on robust statistics,” *IEEE Electronics Letters*, vol. 38, pp. 1233–1234, September 2002.
- [36] X. Zhao, D. O’Shaughnessy, and N. Minh-Quang, “A processing method for pitch smoothing based on autocorrelation and cepstral f0 detection approaches,” in *Signals, Systems and Electronics (ISSSE ’07)*, August 2007, pp. 59–62.
- [37] S. Marchand, “An efficient pitch-tracking algorithm using a combination of fourier transforms,” in *DAFX-01*, Limerick, Ireland, August 2001, pp. 59–62.
- [38] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spo-

- ken word recognition,” *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-26, no. 1, pp. 43–49, February 1978.
- [39] H. Silverman and D. Morgan, “The application of dynamic programming to connected speech recognition,” *ASSP Magazine, IEEE*, vol. 7, no. 3, pp. 6–25, 1990.
- [40] E. Keogh and M. Pazzani, “Scaling up dynamic time warping to massive datasets,” in *3rd European Conference on Principles and Practice of Knowledge Discovery in Databases PKDD’99*, Prague, Czech Republic, September 1999, pp. 1–11.
- [41] A. Shankera and A. Rajagopalan, “Off-line signature verification using dtw,” *Pattern Recognition Letters*, vol. 28, no. 12, pp. 1407–1414, September 2007.
- [42] B. Legrand, C. Chang, S. Ong, S. Neo, and N. Palanisamy, “Chromosome term classification using dynamic time warping,” *Pattern Recognition Letters*, vol. 29, no. 3, pp. 215–222, February 2008.
- [43] V. Tuzcu and S. Nas, “Dynamic time warping as a novel tool in pattern recognition of ecg changes in heart rhythm disturbances,” in *International Conference on Systems, Man and Cybernetics*, Hawaii, USA, October 2005, pp. 182–186.
- [44] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*. New York, NY, USA: Springer Verlag, 2005.
- [45] M. Gubian, F. Torreira, H. Strik, and L. Boves, “Functional data analysis as a

- tool for analyzing speech dynamics. a case study on the french word c'était," in *Interspeech 2009*, Brighton, UK, September 2009, pp. 2199–2202.
- [46] M. Zellers, M. Gubian, and B. Post, "Redescribing intonational categories with functional data analysis," in *Interspeech 2011*, Makuhari, Japan, September 2010, pp. 1141–1144.
- [47] C. Cheng, Y. Xu, and M. Gubian, "Exploring the mechanism of tonal contraction in taiwan mandarin," in *Interspeech 2010*, Makuhari, Japan, 2010.
- [48] F. Ehsani and E. Knodt, "Speech technology in computer-aided language learning: Strengths and limitations of a new call paradigm," *Language Learning & Technology*, vol. 2, no. 1, pp. 613–625, July 1998.
- [49] L. Gu and J. Harris, "Slap: a system for the detection and correction of pronunciation for second language acquisition," in *International Symposium on Circuits and Systems, ISCAS '03*, vol. 2, 2003.
- [50] S. Hamid and M. Rashwan, "Automatic generation of hypotheses for automatic diagnosis of pronunciation errors," in *NEMLAR Conf. on Arabic Language Resources and Tools*, Cairo, Egypt, 2004.
- [51] S. Abdou, S. Hamid, M. Rashwan, A. Samir, O. Abdel-Hamid, M. Shahin, and W. Nazih, "Computer aided pronunciation learning system using speech recognition techniques," in *InterSpeech 2006 ICSLP*, Pittsburgh, PA, USA, September 2006.

- [52] N. Moustoufas and V. Digalakis, “Automatic pronunciation evaluation of foreign speakers using unknown text,” *Comput. Speech Language*, vol. 21, no. 1, pp. 219–230, 2007.
- [53] C. Molina, N. Yoma, J. Wuth, and H. Vivanco, “Asr based pronunciation evaluation with automatically generated competing vocabulary,” *Speech Communication*, vol. 51, no. 6, pp. 485–498, June 2009.
- [54] B. Dong and Y. Yan, “A synchronous method for automatic scoring of language learning,” in *6th International Symposium on Chinese Spoken Language Processing. ISCSLP '08*, 2008.
- [55] J. Tepperman and S. Narayanan, “Using articulatory representations to detect segmental errors in nonnative pronunciation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 8–22, January 2007.
- [56] F. Stouten and J. Martens, “On the use of phonological features for pronunciation scoring,” in *ICASSP, 2006*, 2006.
- [57] P. Su, Q. Chen, and X. Wang, “A fuzzy pronunciation evaluation model for english learning,” in *Fifth International Conference on Machine Learning and Cybernetics*, Dalian, China, August 2006.
- [58] L. Ooppelstrup, M. Blomberg, and D. Elenius, “Scoring children’s foreign language pronunciation,” in *FONETIK*, Gothenburg, Sweden, May 2005.

- [59] J. Bernstein, Cohen, M., Murveit, H., Rtischev, D., Weintraub, and M., “Automatic evaluation and training in english pronunciation,” in *Proc. Int. Congress on Spoken Language Processing ICSLP '90.*, 1990.
- [60] M. Eskenazi, “Detection of foreign speakers’pronunciation errors for second language training - preliminary results,” in *Proc. Proc. Int. Congress on Spoken Language Processing (ICSLP) '96*, 1996.
- [61] S. Hiller, E. Rooney, R. Vaughan, M. Eckert, J. Laver, and M. Jack, “An automated system for computer-aided pronunciation learning,” *Computer Assisted Language Learning*, vol. 7, pp. 51–63, 1994.
- [62] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, “Automatic text-independent pronunciation scoring of foreign language student speech,” in *ICSLP'96*, Philadelphia, PA, USA, October 1996.
- [63] B. Dong, Q. Zhao, J. Zhang, and Y. Yan, “Automatic assessment of pronunciation quality,” in *International Symposium on Chinese Spoken Language Processing*, December 2004.
- [64] J. Tepperman and S. Narayanan, “Better nonnative intonation scores through prosodic theory,” in *InterSpeech 2008*, Brisbane, Australia, September 2008.
- [65] M. Eskenazi and S. Hansma, “The fluency pronunciation trainer,” in *Proc. STiLL Workshop on Speech Technology in Language Learning, Marhollmen*, May 1998.

- [66] M. Peabody and S. Seneff, "Towards automatic tone correction in non-native mandarin," in *Proc. 5th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Kent Ridge, Singapore, December 2006.
- [67] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *ICASSP'97*, vol. 2, 1997.
- [68] R. Delmonte, M. Peterea, and C. Bacalu, "Slim: Prosodic module for learning activities in a foreign language," in *Proc. ESCA, Eurospeech 97*, vol. 2, Rhodes, Greece, 1997.
- [69] H. Kim and W. Sung, "Implementation of an intonational quality assessment system," in *ICSLP-2002*, 1225-1228 2002.
- [70] K. You, H. Kim, and W. Sung, "Implementation of an intonational quality assessment system for a handheld device," in *INTERSPEECH-2004*, 2004.
- [71] J. van Santen, E. Prud'hommeaux, and L. Black, "Automated measures for assessment of prosody," *Speech Communication*, vol. 51, no. 11, pp. 1082–1097, November 2009.
- [72] Y. Kachru, "Discourse strategies, pragmatics and esl. where are we going?" *RELIC Journal*, vol. 16, no. 2, pp. 1–17, 1985.
- [73] M. Celce-Murcia and E. Olshtain, *Discourse and Context in Language Teaching: A Guide for Language Teachers*. Cambridge: Cambridge University Press, 2000.

- [74] C. Busso and S. Narayanan, “Interrelation between speech and facial gestures in emotional utterances: a single subject study,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2331–2347, November 2007.
- [75] P. Ekman and E. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS)*. New York, NY, USA: Oxford University Press, 1997.
- [76] S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, “An acoustic study of emotions expressed in speech,” in *8th International Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea, October 2004, pp. 2193–2196.
- [77] C. Busso, S. Lee, and S. Narayanan, “Analysis of emotionally salient aspects of fundamental frequency for emotion detection,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009.
- [78] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, “The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals,” in *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, August 2007, pp. 2253–2256.
- [79] I. Fonagy, “A new method of investigating the perception of prosodic features,” *Language and Speech*, vol. 21, no. 1, pp. 34–49, 1978.

- [80] R. van Bezooijen, *The characteristics and recognizability of vocal expression of emotions*. Dordrecht, The Netherlands: Foris, 1968.
- [81] M. Goudbeek, J. Goldman, and K. Scherer, “Emotion dimensions and formant position,” in *Interspeech 2009*, Brighton, UK, September 2009, pp. 1575–1578.
- [82] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve, “Approaching automatic recognition of emotion from voice: A rough benchmark,” in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Newcastle, Northern Ireland, UK: ISCA, September 2000, pp. 207–212.
- [83] F. Dellaert and T. P. A. Waibel, “Recognizing emotion in speech,” in *International Conference on Spoken Language (ICSLP 1996)*, vol. 3, Philadelphia, PA, USA, October 1996, pp. 1970–1973.
- [84] M. Sedaaghi, C. Kotropoulos, and D. Ververidis, “Using adaptive genetic algorithms to improve speech emotion recognition,” in *International Workshop on Multimedia Signal Processing (MMSP 2007)*, Chania, Crete, Greece, October 2007, pp. 461–464.
- [85] A. Alvarez, I. Cearreta, J. López, A. Arruti, E. Lazkano, B. Sierra, and N. Garay, “Feature subset selection based on evolutionary algorithms for automatic emotion recognition in spoken Spanish and standard Basque language,” in *Ninth International Conference on Text, Speech and Dialogue (TSD 2006)*, Brno, Czech Republic, September 2006, pp. 565–572.

- [86] D. Patterson and D. Ladd, “Pitch range modelling: Linguistic dimensions of variation,” in *Proceedings of the XIVth International Congress of Phonetic Sciences (ICPhS 1999)*, San Francisco, CA, USA, August 1999, pp. 1169–1172.
- [87] P. Lieberman and S. Michaels, “Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech,” *Journal of the Acoustical Society of America*, vol. 34, no. 7, pp. 922–927, July 1962.
- [88] A. Paeschke, “Global trend of fundamental frequency in emotional speech,” in *Speech Prosody (SP 2004)*, Nara, Japan, March 2004, pp. 671–674.
- [89] L. Yang and N. Campbell, “Linking form to meaning: the expression and recognition of emotions through prosody,” in *ISCA ITRW on Speech Synthesis*, Perthshire, Scotland, August-September 2001.
- [90] M. Rotaru and D. J. Litman, “Using word-level pitch features to better predict student emotions during spoken tutoring dialogues,” in *9th European Conference on Speech Communication and Technology (Interspeech’2005 - Eurospeech)*, Lisbon, Portugal, September 2005, pp. 881–884.
- [91] E. Grabe, G. Kochanski, and J. Coleman, “Connecting intonation labels to mathematical descriptions of fundamental frequency,” *Language and Speech*, vol. 50, no. 3, pp. 281–310, October 2007.
- [92] C. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, “Emotion recognition based on phoneme classes,” in *8th Inter-*

- national Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea, October 2004, pp. 889–892.
- [93] B. Schuller, R. Müller, M. Lang, and G. Rigoll, “Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles,” in *9th European Conference on Speech Communication and Technology (Interspeech’2005 - Eurospeech)*, Lisbon, Portugal, September 2005, pp. 805–808.
- [94] D. Bitouk, R. Verma, and A. Nenkova, “Class-level spectral features for emotion recognition,” *Speech Communication*, vol. 52, no. 1, pp. 613–625, 2010.
- [95] C. Busso, S. Lee, and S. Narayanan, “Using neutral speech models for emotional speech analysis,” in *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, August 2007, pp. 2225–2228.
- [96] C. Busso and S. Narayanan, “Joint analysis of the emotional fingerprint in the face and speech: A single subject study,” in *International Workshop on Multimedia Signal Processing (MMSP 2007)*, Chania, Crete, Greece, October 2007, pp. 43–47.
- [97] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schroder, “Feeltrace: An instrument for recording perceived emotion in real time,” in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Ireland, UK, September 2000, pp. 19–24.

- [98] P. Traynor, “Effects of computer-assisted-instruction on different learners,” *Instructional psychology journal*, pp. 137–143, 2003.
- [99] E. Bernat, “Assessing eap learners’beliefs about language learning in the australian context,” *Asian EFL Journal*, vol. 8, no. 2, June 2006.
- [100] A. Moyer, *Age, accent and experience in second language acquisition: An integrated approach to critical period inquiry*. Clevedon: Multilingual Matters, 2004.
- [101] H. Baetens, *Bilingualism: basic principles*. On-line version, 1982.
- [102] F. Saussure, S. Bouquet, and R. Engler, *Writings in general linguistics*. Oxford: Oxford University Press, 2006.
- [103] J. Holmes and W. Holmes, *Speech synthesis and recognition*. 2nd edition. CRC Press, 2001.
- [104] and El-Imam Y.A. and D. Z.M., “Rules and algorithms for phonetic transcription of standard malay,” *IEICE Transactions on Information and Systems*, 2005.
- [105] M. Rypa and P. Price, “Vilts: A tale of two technologies,” *CALICO Journal*, vol. 16, no. 3, pp. 385–404, 1999.
- [106] M. Shimizu and M. Taniguchi, “Reaffirming the effect of interactive visual feedback on teaching english intonation to japanese learners,” in *Phonetics Teaching and Learning Conference 2005, University College London*, 2005.

- [107] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and K. Sonmez, “Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners,” in *ICSLP, 2000*, 2000.
- [108] D. Jurafsky and J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Second Edition, January, 2009, 2009.
- [109] H. Jia, J. Tao, and X. Wang, “Prosody variation: Application to automatic prosody evaluation of mandarin speech,” in *Speech Prosody 2008*, 2008.
- [110] J. Jenkins, *The Phonology of English as an international language*. Oxford University Press, 2000.
- [111] D. Bolinger, *Intonation and its parts: Melody in spoken English*. Stanford: Stanford University Press., 1986.
- [112] I. Fónagy, *Languages within Language: An Evolutive Approach*. Amsterdam/Philadelphia: John Benjamins, 2001.
- [113] N. Bell, “Responses to failed humor,” *Journal of Pragmatics*, vol. 41, no. 9, pp. 1825–1836, September 2008.
- [114] D. Bolinger, *Intonation and its uses: melody in grammar and discourse*. Stanford: Stanford University Press., 1989.

- [115] A. Peters, “Language learning strategies: does the whole equal the sum of the parts?” *Lang*, vol. 53, pp. 560–73, 1977.
- [116] E. Grabe and B. Post, “Intonational variation in the british isles,” 2002.
- [117] T. Face, “Narrow focus intonation in castillian spanish absolute negatives,” *Journal of Language and Linguistics*, vol. 5, no. 2, pp. 295–311, 2006.
- [118] D. Ramírez and J. Romero, “The pragmatic function of intonation in l2 discourse: English tag questions used by spanish speakers,” *Intercultural Pragmatics*, vol. 2, pp. 151–168, June 2005.
- [119] J. Morley, “The pronunciation component in teaching english to speakers of other languages,” *TESOL Quarterly*, vol. 25, no. 3, pp. 481–520, 1991.
- [120] M. Raman, *English Language Teaching*. New Delhi: Atlantic Publishers & Distributors, 2004.
- [121] M. Pennington, “Teaching pronunciation from the top down.” *RELC Journal*, vol. 20, no. 1, pp. 20–38, 1989.
- [122] C. Dalton and B. Seidlhofer, *Pronunciation*. Oxford: Oxford University Press, 1994.
- [123] R. Carter and D. Nunan, *The Cambridge Guide to Teaching English to Speakers of Other Languages*. Cambridge: Cambridge University Press, 2001.

- [124] R. Jones, “Beyond ’listen and repeat’: Pronunciation teaching materials and theories of second language acquisition.” *System* 25, vol. 1, pp. 103–112, 1997.
- [125] L. Rabiner, “On creating reference templates for speaker independent recognition of isolated words,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 34–42, 1978.
- [126] L. Rabiner and J. Wilpon, “Speaker-independent isolated word recognition for a moderate size (54) word vocabulary,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 6, pp. 583–587, 1979.
- [127] L. Rabiner and C. Schmidt, “Application of dynamic time warping to connected digit recognition,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 377–388, 1980.
- [128] S. Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.
- [129] H. Tao, *Units in Mandarin Conversion: Prosody, Discourse & Grammar*. John Benjamins, 1996.
- [130] P. Boersma and D. Weeninck, “Praat, a system for doing phonetics by computer,” Institute of Phonetic Sciences of the University of Amsterdam, Amsterdam, Netherlands, Technical Report 132, 1996, <http://www.praat.org>.

- [131] L. Gillick and S. Cox, “Some statistical issues in the comparison of speech recognition algorithms.” in *ICASSP’89*, Glasgow, Scotland, 1989.
- [132] T. Bänziger and K. Scherer, “The role of intonation in emotional expressions,” *Speech Communication*, vol. 46, no. 3-4, pp. 252–267, July 2005.
- [133] C. Busso, M. Bulut, and S. Narayanan, “Toward effective automatic recognition systems of emotion in speech,” in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, S. M. J. Gratch, Ed. New York, NY, USA: Oxford University Press, 2010.
- [134] H. Wang, A. Li, and Q. Fang, “F0 contour of prosodic word in happy speech of Mandarin,” in *Affective Computing and Intelligent Interaction (ACII 2005)*, *Lecture Notes in Artificial Intelligence 3784*, J. Tao, T. Tan, and R. Picard, Eds. Berlin, Germany: Springer-Verlag Press, November 2005, pp. 433–440.
- [135] G. McKeown, M. Valstar, R. Cowie, and M. Pantic, “The SEMAINE corpus of emotionally coloured character interactions,” in *IEEE International Conference on Multimedia and Expo (ICME 2010)*, Singapore, July 2010, pp. 1079–1084.
- [136] J. Arias, N. Yoma, and H. Vivanco, “Automatic intonation assessment for computer aided language learning,” *Speech Communication*, vol. 52, no. 1, pp. 254–267, March 2010.
- [137] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” *IEEE Transactions on*

- Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, January 2009.
- [138] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,” *Speech Communication*, vol. 53, no. 1, pp. 1062–1087, December 2011.
- [139] A. Paeschke and W. Sendlmeier, “Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements,” in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Ireland, UK, September 2000, pp. 75–80.
- [140] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, “ToBI: a standard for labelling english prosody,” in *2th International Conference on Spoken Language Processing (ICSLP 1992)*, Banff, Alberta, Canada, October 1992, pp. 867–870.
- [141] J. Liscombe, J. Venditti, and J. Hirschberg, “Classifying subject ratings of emotional speech using acoustic features,” in *8th European Conference on Speech Communication and Technology (EUROSPEECH 2003)*, Geneva, Switzerland, September 2003, pp. 725–728.
- [142] P. Taylor, “Analysis and synthesis of intonation using the tilt model,” *Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1697–1714, March 2000.
- [143] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, “An articulatory study

- of emotional speech production,” in *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, Lisbon, Portugal, September 2005, pp. 497–500.
- [144] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, “Primitives-based evaluation and estimation of emotions in speech,” *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, October-November 2007.
- [145] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of German emotional speech,” in *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, Lisbon, Portugal, September 2005, pp. 1517–1520.
- [146] A. Batliner, D. Seppi, S. Steidl, and B. Schuller, “Segmenting into adequate units for automatic recognition of emotion-related episodes: A speech-based approach,” *Advances in Human-Computer Interaction*, pp. 1–15, 2010.
- [147] B. Vlasenko, D. Prylipko, D. Philippou-Hübner, and A. Wendemuth, “Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions.” in *12th Annual Conference of the International Speech Communication Association (Interspeech'2011)*, Florence, Italy, August 2011, pp. 1577–1580.
- [148] F. Eyben, M. Wöllmer, and B. Schuller, “openEAR-introducing the munich open-source emotion and affect recognition toolkit,” in *International Conference on*

- Affective Computing and Intelligent Interaction (ACII 2009)*, Amsterdam, The Netherlands, September 2009, pp. 576–581.
- [149] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. Narayanan, “The INTERSPEECH 2010 paralinguistic challenge,” in *Interspeech 2011*, Makuhari, Japan, September 2010, pp. 2794–2797.
- [150] J. Jeon, R.X., and Y. Liu, “Sentence level emotion recognition based on decisions from subsentence segments,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May 2011.
- [151] T. Kudoh and Y. Matsumoto, “Use of support vector learning for chunk identification,” in *Workshop on Learning language in logic (LLL 2000) and conference on Computational natural language learning (CoNLL 2000)*, Lisbon, Portugal, September 2000.
- [152] D. Paul and J. Baker, “The design for the Wall Street Journal-based CSR corpus,” in *2th International Conference on Spoken Language Processing (ICSLP 1992)*, Banff, Alberta, Canada, October 1992, pp. 899–902.
- [153] J. Kittler, “Combining classifiers: A theoretical framework,” *Pattern Analysis & Applications*, vol. 1, no. 1, pp. 18–27, March 1998.
- [154] J. Kim, H. Rao, and M. Clements, “Investigating the use of formant based features for detection of affective dimensions in speech,” in *Affective Computing and Intelligent Interaction (ACII 2011)*, ser. *Lecture Notes in Computer Science*, S.

- DMello, A. Graesser, B. Schuller, and J.C. Martin, Eds. Memphis, TN, USA: Springer Berlin / Heidelberg, October 2011, pp. 369–377.*
- [155] H. Meng and N. Bianchi-Berthouze, “Naturalistic affective expression classification by a multi-stage approach based on hidden markov models,” in *Affective Computing and Intelligent Interaction (ACII 2011)*, ser. *Lecture Notes in Computer Science*, S. DMello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Memphis, TN, USA: Springer Berlin / Heidelberg, October 2011, pp. 378–387.
- [156] L. Cen, Z. Yu, and M. Dong, “Speech emotion recognition system based on l1 regularized linear regression and decision fusion,” in *Affective Computing and Intelligent Interaction (ACII 2011)*, ser. *Lecture Notes in Computer Science*, S. DMello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Memphis, TN, USA: Springer Berlin / Heidelberg, October 2011, pp. 332–340.
- [157] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, “On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues,” *Journal on Multimodal User Interfaces*, vol. 3, pp. 7–19, March 2010.
- [158] M. Nicolaou, H. Gunes, and M. Pantic, “Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space,” *IEEE Transactions on Affective Computing*, vol. 2, pp. 92–105, April-June 2011.
- [159] J. Russell and B. Fehr, “Relativity in the perception of emotion in facial ex-

pressions,” *Journal of Experimental Psychology: General*, vol. 116, pp. 223–237,  
September 1987.

# Publicaciones del autor

## Artículos ISI como primer autor

1. **Arias, J.P.**, , Busso, C., Yoma, N.B., (2012), “*Shape-based modeling of the fundamental frequency contour for emotion detection in speech,*” Submitted to Computer Speech and Language (Elsevier).
2. **Arias, J.P.**, Yoma, N.B., Vivanco, H., (2010), “*Automatic intonation assessment for computer aided language learning,*” Speech Communication (Elsevier), Volume 52, Issue 3, March 2010, pp. 254-267.

## Presentaciones en congresos como primer autor

1. **Arias, J.P.**, Yoma, N.B., Vivanco, H., (2009), “*Word stress assessment for computer aided language learning,*” in Proceedings of Interspeech 2009, 6-10 September, Brighton, UK.