CARACTERIZACIÓN ACÚSTICA DEL ACENTO BASADA EN CORPUS: UN ENFOQUE MULTILINGÜE INGLÉS/ESPAÑOL

David Escudero Mancebo[1], Lourdes Aguilar[2], César González Ferreras[1], Carlos Vivaracho Pascual[1], and Valentín Cardeñoso-Payo[1]

[1] Departamento de Informática, Universidad de Valladolid, España
[2] Departamento de Filología Hispanica, Universidad Autónoma de Barcelona, España

[1] {descuder,cesargf,cevp, valen}@infor.uva.es
[2] {Lourdes.aguilar}@uab.es

Resumen

El objetivo del artículo es identificar las propiedades prosódicas más relevantes de cara a la identificación del acento de prominencia en un contexto multilingüe usando un clasificador de tipo red neuronal y un árbol de decisión. El enfoque multilingüe consiste en utilizar datos que proceden del análisis de corpus de hablada en inglés para su aplicación en el español. Como principal resultado, se comprueba que el enfoque multilingüe permite obtener tasas elevadas de identificación del acento en corpus de habla, aunque no deja de presentar los mismos problemas que el enfoque multilocutor.
Palabras Clave: Caracterización prosódica, análisis de corpus, prosodia y variación lingüística

Abstract

In this work we present an experiment of characterization of the accent in multilingual contexts. By characterizing the accent, we understand the analysis of the most relevant prosodic features facing the problem of the accent identification. The accent or prominence is the prosodic function that makes relevant a given unit of the message with respect to the other units of the message (in this work the unit is lexical word). A neural network and a decision tree are trained with different corpora so that the relevant prosodic features are identified. The multilingual approach consists of using data both in English and Spanish. Our main result is the multilingual approach presents the same problems as the multispeaker does, but, in spite of it, it permits to obtain high identification rates of the accent in the corpora.
Key words: Prosodic representation, analysis of corpora, crossligual prosody.

1.- Introducción

El objetivo del artículo es identificar las propiedades prosódicas más relevantes de cara a la caracterización del acento de prominencia, de manera que se puedan aplicar en la identificación automática de corpus extensos de habla. Para entender la prominencia acentual, debemos diferenciar entre el concepto de acento léxico (stress) y el de acento melódico. En español, todas las palabras con contenido léxico acarrean acento léxico, que en general se manifiesta mediante un conjunto de propiedades acústicas: las sílabas tónicas presentan

valores superiores de duración, intensidad y tono superiores de las que las sílabas átonas. Otro tipo de acento es el melódico (accent), y cabe señalar, a este respecto, que no todos los acentos son melódicos en español, pero el hecho de que algunos de ellos lo sean, acarrea información importante sobre la organización prosódica de los enunciados.

El acento tonal o melódico (accent) caracteriza a las sílabas acentuadas cuando aparecen en contextos con prominencia de frase, y se define como una inflexión entonativa local (sea el movimiento ascendente, descendente o complejo) articulada alrededor de las sílabas acentuadas.

En este sentido, la identificación de unidades de prominencia tiene aplicaciones en diversas áreas de las tecnologías del habla. Sankaranarayanan/Narayanan (2008) apuntan beneficios en aplicaciones de reconocimiento del habla, síntesis de voz y sistemas de diálogo. En la bibliografía existente sobre caracterización o identificación automática de prominencias se obtienen tasas de reconocimiento cercanas al 80% (ver (Rangarajan/Narayanan 2008) para una excelente revisión). La mayor parte de las aproximaciones funcionan empleando técnicas de aprendizaje basado en corpus, es decir, los clasificadores (métodos automáticos) se entrenan empleando corpus que están previamente enriquecidos con etiquetas prosódicas. El uso de este tipo de técnicas para el español es, por tanto, limitado dada la escasez de corpus transcritos con marcas prosódicas que describan la entonación del español.

Llegados a este punto, cabe mencionar que el etiquetado prosódico manual de corpus de habla es una tarea compleja que requiere un esfuerzo nada despreciable de equipos de transcriptores expertos y previamente entrenados. Así Syrdal/Hirschberg/Beckman (2001) estiman entre 100 y 200 veces tiempo real el coste necesario para disponer de etiquetado ToBI. No es banal apuntar ahora que una de las aplicaciones principales de la caracterización automática del acento es proporcionar corpus enriquecidos con etiquetas que los transcriptores humanos puedan revisar de forma rápida, reduciendo el tiempo total del trabajo. En este marco de investigación, el equipo que firma este artículo se encuentra actualmente implicado en el proyecto Glissando de creación de un corpus etiquetado prosódicamente para el español y el catalán.

Con el objetivo último de desarrollar herramientas de ayuda a la transcripción prosódica, en este artículo sondeamos la posibilidad de utilizar modelos entrenados con un corpus en inglés para etiquetar corpus en idiomas diferentes, en particular, en español. Los beneficios de emplear este enfoque son evidentes debido a la escasez de material etiquetado prosódicamente especialmente en las lenguas románicas. Sin embargo somos conscientes de los desafíos y en este artículo analizamos algunos de ellos.

Partimos de un enfoque de clasificación multilingüe donde se entrenan clasificadores empleando un corpus hablado en una lengua para etiquetar un corpus en otra lengua diferente. Las diferentes características prosódicas de energía, F0 y duración se contrastan sistemáticamente para analizar las diferencias tanto en rango como en capacidad de las mismas para caracterizar el acento. Por otro lado, analizar con los mismos criterios un corpus multilocutor sirve para investigar si las diferencias entre idiomas pueden aparecer también entre diferentes locutores. El resultado final apunta a que la diferencia entre las características prosódicas que caracterizan el acento de prominencia en los diferentes idiomas es importante pero también lo son las diferencias entre locutores. Las altas tasas de rendimiento en la clasificación de los acentos en escenarios multilocutor validan el procedimiento.

El artículo detalla en primer lugar el procedimiento experimental presentando los corpus empleados y los clasificadores que se utilizan, y a continuación se detallan y discuten los resultados obtenidos.

2 El procedimiento experimental

El enfoque multilingüe consiste en el entrenamiento con los datos de un corpus hablado en un idioma determinado y la realización de las pruebas con los datos de un corpus hablado en otro idioma diferente. Comparamos las diferencias entre idiomas y las diferencias entre locutores utilizando los datos de un corpus que contiene seis locutores diferentes. Por otra parte, se van a contrastar sistemáticamente en varios aspectos prácticos, las diferencias que aparecen entre las características prosódicas en los diversos corpus.

En primer lugar, se analiza la escala de las características prosódicas para contrastar las diferencias entre las lenguas y locutores. El cruce de diferentes locutores de diferentes idiomas en tareas de entrenamiento y prueba, muestra el claro impacto de la variabilidad de escala en el rendimiento de los clasificadores, lo que justifica la necesidad del proceso de normalización.

El estudio multilingüe continúa con el examen de la pertinencia de las propiedades prosódicas para caracterizar los diferentes corpus. Estas características prosódicas se clasifican en términos de su capacidad informativa para discriminar si una palabra o una sílaba es acentuada o si no es. Para cada idioma y locutor se crea un ranking diferente que será contrastado. Además, también se analizan las características más informativas para verificar si aparecen diferencias significativas entre los diferentes corpus.

Finalmente, se contrastan los resultados automáticos de predicción con juicios perceptivos realizados por un conjunto de etiquetadores que utilizaron el mismo corpus de pruebas. Los resultados de esta última prueba muestran la utilidad del proceso de etiquetado automático de cara a su posible aplicación en procesos de etiquetado prosódico manual de corpus.

Corpus	ESMA	BURNC	BURNC.f1a	BURNC.f2b	BURNC.f3a	BURNC.m1b	BURNC.m2b	BURNC.m3b
#palabras	7236	27767	3790	77994	2624	3974	3413	1972
#acentuadas	2483	13899	2053	6214	1284	1604	1823	924
#no-acent.	4895	14586	7831	6057	1438	2467	1700	1093

Tabla 1: Número de palabras en los corpus y subcorpus

2.1 Procesamiento de corpus

Empleamos el corpus Boston University Radio News Corpus BURNC (Ostendorf/Shattuck 1995) para modelar los acentos en inglés. Este corpus contiene etiquetas que separan fonemas, sílabas y palabras. Los acentos tienen una etiqueta ToBI y también se indica su posición. Siguiendo otros trabajos del estado del arte (Ananthakrishnan/Narayanan 2008) y (Rangarajan/Narayanan 2008), los tonos se alinean con respecto a la sílaba prominente y a la palabra que lo contiene. Empleamos todas las noticias de todos los locutores del corpus (mujeres: f1a, f2b, f3a y hombre m1b, m2b y m3b) como se muestra en la Tabla 1.

El corpus español usado en este trabajo es el corpus llamado ESMA-UPC. Este corpus fue diseñado en la Universidad Politécnica de Cataluña (http://www.gps-tsc.upc.es) para la construcción de un sistema de conversión texto voz basado en la técnica de concatenación de

unidades para el catalán y español (Bonafonte/Moreno 2008). El corpus contiene grabaciones de cerca de tres horas de locuciones habladas en los dos idiomas. A pesar de que el corpus no fue diseñado específicamente para los estudios de la prosodia, contiene datos suficientes para permitir obtener resultados significativos en este trabajo. El corpus fue adquirido en condiciones de estudio de grabación en dos canales separados a 32 kHz. La voz fue grabada en uno de los canales y la salida de un laringógrafo en el otro. Los datos fueron etiquetados de forma automática y revisados manualmente. El etiquetado incluye silencios, la transcripción alofónica, y los límites alofónicos. Esta información se ha incrementado incluyendo los límites entre sílabas, límites entre palabras y las posiciones de acento. La Tabla 1 resume las cantidades de este corpus.

La extracción de características para cada uno de los dos corpus se llevó a cabo siguiendo experimentos similares encontrados en la bibliografía (Ananthakrishnan/Narayanan 2008). Las características prosódicas que se emplean son para la frecuencia: rango de F0 dentro de la palabra (*f0_range*), la diferencia entre el máximo y promedio dentro de la palabra (*f0_maxavg_diff*), la diferencia entre el promedio y el mínimo en la palabra (*f0_minavg_diff*), La diferencien entre el promedio de F0 dentro de la palabra y la media en la frase (*f0_avgutt _diff*), para la energía: el rango de energía dentro de la palabra (*e_range*), la diferencia entre el máximo y el promedio de la energía dentro de la palabra (*e_maxavg_diff*), la diferencia entre el promedio y el mínimo de energía dentro de la palabra (*e_minavg_diff*), y para la duración: duración máxima de todas las vocales dentro de la palabra (duraciones normalización en función del tipo vocal) (*duration*).

La información sintáctico-léxica relacionada con el etiquetado POS ha demostrado ser útil en la mejora de los resultados de clasificación de acentos (Ananthakrishnan/Narayanan 2008) y (Rangarajan/Narayanan 2008). El problema ahora es que no existe una correspondencia evidente entre las etiquetas POS que se utilizan en cada corpus: el corpus BURNC utiliza las etiquetas Penn Treebank (calculadas con el etiquetador BBN (Meteer/Schwartz/Weischedel 1991)) y el corpus ESMA usa EAGLES (obtenidas con el etiquetador Freeling http://nlp.lsi.upc.edu/freeling). Aquí se decidió utilizar la clasificación clásica que asigna a las diferentes palabras del enunciado el papel de palabra función frente al rol de palabra contenido. Esta clasificación se utiliza ampliamente en el modelado de la entonación del español en aplicaciones de síntesis texto a voz (Escudero/Cardeñoso/Valentín 2007). Las etiquetas Penn Treebank se han agrupado de manera que las palabras función fueron: EX (existencial), RP (de partículas), CC (conjunción de coordinación), DT (determinante), IN (preposición, conjunción subordinante), WDT (WH-determinante, A (preposición) y CD (número cardinal). El resto de los tipos de palabras se consideran palabras contenido. Las palabras del corpus ESMA están previamente clasificadas en base a este criterio porque este corpus incluye etiquetas de grupo acentual (un grupo formado por una palabra contenido precedida por una serie de palabras función).

En cuanto al contexto, nos centramos en los efectos locales (a nivel de palabra sílaba) porque el contexto puede ser altamente dependiente de la lengua y el modelado de la de sus efectos está fuera del alcance de este artículo.

	ESMA (Español)		Boston Corpus (Inglés)											
			F1a		F2b		F3a		M1b		M2b		M3b	
Propiedad	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
f0_range	48.2	26.3	39.6	39.7	56.0	43.3	42.2	42.9	26.7	30.6	24.7	30.7	28.0	27.3
f0_maxavg_diff	22.8	16.1	19.1	21.2	25.1	21.7	18.7	21.0	14.5	19.5	12.2	16.8	13.2	13.5
F0_minavg_diff	25.4	14.6	20.4	21.8	30.9	26.4	23.5	25.8	12.3	13.8	12.4	16.0	14.7	15.4
F0_avgutt_diff	-0.8	18.4	-19	57.5	-5.5	42.5	-21	62.8	-13	40.5	-28	60.2	-15	45.7

e_range	18.6	8.5	13.9	6.8	16.7	6.4	13.7	6.3	12.9	6.5	12.4	6.9	11.5	5.2
e_maxavg_diff	10.0	5.4	7.7	4.4	9.2	4.1	7.7	4.0	7.7	4.2	7.8	4.8	6.9	3.6
e_minavg_diff	8.6	4.1	6.2	3.1	7.6	3.4	6.0	3.1	5.2	3.0	4.7	2.8	4.6	2.2
duration	-0.9	9.3	2.5	9.8	4.2	10.6	1.4	12.0	1.0	9.9	-0.5	12.0	1.2	9.5

Tabla 2: Estadísticas de las características de los diferentes corpus y subcorpus. Unidades: f0_range, f0_maxavg_diff, f0_minavg_diff y de f0_avgutt_diff en Hz, e_range e_maxavg_diff y e_minavg_dif en RMSE/100, y la duración se normaliza * 10

2.2 Los clasificadores

Se empleó el paquete software de aprendizaje Weka (Hall/Eibe/Pfahringer/Reutemann/Witten 2009) para construir los árboles de decisión C4.5 (J48 en Weka). Se emplearon diferentes valores para el umbral de confianza de la poda del árbol, aunque los mejores resultados se obtienen con el valor por defecto (0,25). El número mínimo de casos por hoja también se establece en el valor por defecto (2).

Se emplea también un perceptrón multicapa (MLP) no lineal que utiliza el algoritmo de aprendizaje de retropropagación de errores. Se pusieron a prueba varias configuraciones de la red, logrando los mejores resultados con la siguiente: i) una sola capa oculta con 12 neuronas, ii) 100 épocas de entrenamiento, iii) dos neuronas en la capa de salida, una para cada clase se clasifica.

Debido a la escala diferente de las características de los corpus de entrenamiento, se han probado diferentes técnicas de normalización: la división Z-Norm, Min-Max, división por el máximo y la norma euclídea. Finalmente la normalización se ha procesado por corpus y por locutor mediante la técnica de Z-Norm. En (González/Vivaracho/Escudero/Cardeñoso 2010) se muestra el impacto negativo del desequilibrio de los datos de entrada sobre el resultado final. Para evitarlo se aplican métodos de re-muestreo como son la repetición de la clase minoritaria (Vivaracho/Simon 2010) para el clasificador MLP y la técnica SMOTE (Chawla/Kevin/Kegelmeyer 2002) para el clasificador C4.5.

	ESMA-UPC	BURNC	BURNC. f1a	BURNC. f2b	BURNC. f3a	BURNC. m1b	BURNC. m2b	BURNC. m3b
ESMA-UPC	**86.6/81.0**	72.7/76.5	75.6/76.0	74.7/76.1	76.5/77.9	82.7/76.0	73.6/75.5	75.9/74.7
BURNC	81.4/60.3	**80.5/80.4**	--	--	--	--	--	--
BURNC.f1a	71.1/72.1	--	**83.2/80.3**	79.8/78.3	76.9/74.6	78.7/77.4	80.6/80.4	78.0/76.7
BURNC.f2b	81.5/65.5	--	81.5/80.0	**84.6/82.9**	78.6/74.3	79.0/72.6	81.6/74.5	79.4/75.1
BURNC.f3a	80.9/78.6	--	80.7/79.5	79.0/79.8	**82.2/80.3**	80.3/77.8	82.4/81.6	79.1/77.3
BURNC.m1b	76.6/75.9	--	77.6/77.0	78.0/76.8	76.7/75.6	**84.7/80.8**	74.7/76.9	77.8/75.0
BURNC.m2b	74.4/63.0	--	80.5/79.5	77.8/75.1	78.3/73.5	79.1/74.4	**83.0/82.3**	78.1/75.8
BURNC.m3b	69.3/75.5	--	81.5/80.8	78.4/78.9	78.1/77.2	79.9/78.6	79.9/81.0	**81.0/76.6**

Tabla 3: Tasas de clasificación (en porcentajes) cuando se usan palabras como unidad de análisis considerando la presencia de acento. El corpus de entrenamiento aparece en las filas, y el de prueba en las columnas. En las celdas (xx / yy), donde xx es el porcentaje de clasificación obtenidos con el clasificador C4.5, yy con el clasificador MLP

3 Resultados

La Tabla 2 muestra los valores promedio y las desviaciones estándar de las propiedades prosódicas de los distintos corpus y subcorpus analizados en este trabajo. Las diferencias entre los locutores masculinos y femeninos, en lo que hace referencia a los valores de F0, son claramente observables (La gama de valores μ de F0 van desde 24.7Hz a 28.0Hz para hablantes masculinos, mientras que para los hablantes femeninos van desde 42.2Hz a 56.0Hz). Los valores de F0 parecen ser más estable en el corpus de la ESMA (Los valores de

σ van desde 14.6Hz a 26.3Hz) que en los subcorpus de BURNC (σ de 13.5Hz a 62.8Hz). En el caso de las variables relacionadas con la energía, también hay diferencias significativas entre los corpus. El corpus BURNC parece ser más estable σ va desde 2,2 hasta 6,9 RMSE/100, frente a la variabilidad observada en el corpus de la ESMA que va desde 4,1 hasta 8,5 RMSE/100. La variable de duración también muestra diferencias significativas entre los diversos corpus.

La Tabla 3 muestra las tasas de clasificación que se logran cuando los distintos corpus intercambian su función de entrenamiento y prueba. En los escenarios convencionales (mismo corpus para entrenamiento y pruebas; diagonal de la Tabla 3), los resultados van desde 80,5 a 86,6%, que son los esperados de acuerdo con el estado del arte: (Rangarajan/ Narayanan 2008) informa de resultados que van desde el 75,0% al 87,7% con el Boston Radio Corpus con palabras como unidad de referencia básica. En el escenario de igual idioma, mismo locutor se obtienen los mejores resultados (celdas de la diagonal en la tabla 3). Las tasas de clasificación disminuyen fuera de dicha diagonal y son altamente dependientes de los subcorpus utilizados. El mejor y peor resultados son 82,7% y 69,3% en el escenario multilingüe y 82,4% y un 74,7% en el escenario inter-locutores. Todos estos porcentajes se refieren al uso de árboles de decisión que parecen ser más eficaces que las redes neuronales.

La Tabla 4 compara la ganancia de información de las distintas características analizadas, proporcionando un medida de la posible pérdida de entropía que se generaría si la división del conjunto de entrenamiento se llevara a cabo en términos dicha característica (Witten/Frank 1999). El etiquetado del corpus de español parece basarse principalmente en características de F0, porque las cuatro características más relevantes están relacionadas con F0 (con excepción de POS) y la diferencia con respecto a las características de energía y duración es muy importante. El etiquetado del corpus en inglés también parece basarse principalmente en las características F0 (*f0_minavg_diff* y *f0_range* comparten la primera posición ranking en los dos corpus). Sin embargo, la energía y la duración parecen ser más relevantes para los transcriptores de inglés para los de español. Este comportamiento parece ser dependiente de los locutores analizados: por ejemplo m3b da más importancia a la energía que f3a. Se han seleccionados los locutores f3a y m3b en este análisis por ser la mejor y peor predicción de los acentos españoles con el árbol de decisión C4.5 tal y como se observa en la Tabla 3. La POS aparece como una de las características más informativos en todos los casos (Esta propiedad no está en un puesto muy alto del ranking para el corpus BURNC.m3b pero tiene un valor alto de IG).

ESMA-UPC		BURNC.f2b		BURNC.m3b	
Caracteristica	IG	Característica	IG	Característica	IG
f0_minavg_diff	0.18888	f0_minavg_diff	0.232	F0_minavg_diff	0.245
f0_range	0.18246	f0_range	0.214	F0_range	0.232
pos	0.17347	pos	0.199	F0_maxavg_diff	0.206
f0_avgutt_diff	0.15215	duration	0.177	E_range	
f0_maxavg_diff	0.10891	f0_maxavg_diff	0.156	E_maxavg_diff	0.165
e_range	0.09695	e_range	0.152	pos	0.164
e_minavg_diff	0.08156	e_maxavg_diff	0.13	E_minavg_diff	0.15
e_maxavg_diff	0.07681	f0_avgutt_diff	0.12	duration	0.139
duration	0.0063	e_minavg_diff	0.105	f0_avgutt_diff	0.117

Tabla 4: Ganancia de Información (IG del término en inglés *Information Gain*), calculada con el software WEKA, de las características que se utilizan para clasificar los acentos en los diferentes corpus.

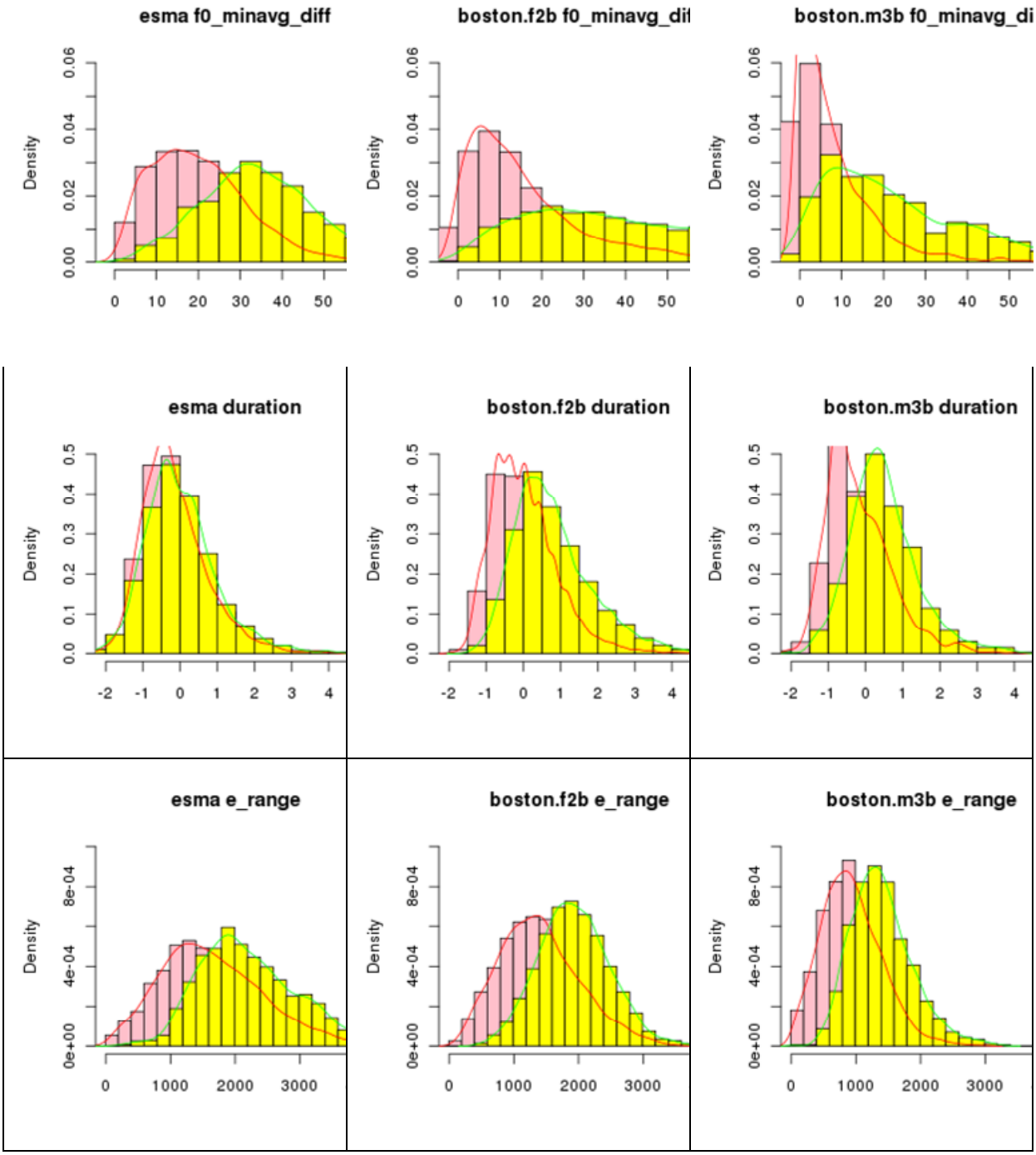

Figura 1: Pares de distribuciones que discriminan palabras acentuadas (en amarillo) frente a palabras no acentuadas (en rosa). En cada fila se analiza una variable distinta: de arriba abajo *f0_minavg_diff, duration* y *e_range*. En cada columna se analiza un subcorpus: de izquierda a derecha: *esma, BURNC* locutor *f2b y BURNC* locutor *m3b.*

La Figura 1 muestra la distribución de las variables *f0_minavg_diff, duration, e_range* para los distintos subcorpus analalizados en la Tabla 4. Se observa que la variable *f0_minavg_diff* discrimina más que las otras dos porque el par de distribuciones en cada gráfica de la fila superior están más separadas de lo que lo están las distribuciones correspondientes a las otras dos variables en las otras dos filas para todos los subcorpus, o lo que es lo mismo para todas

las columnas. La energía (fila inferior) parece ser más discriminante para el inglés, especialmente para el locutor *m3b* de lo que lo es para el español (la celda correspondiente a *esma* y *e_range* nosepara de forma significativa el par de distribuciones). La duración no parece discriminar en absoluto el caso *esma* (distribuciones solapadas en la celda *esma duration*). Estos resultados son coincidentes con los presentados en el párrafo anterior en el que se describe la Tabla 4.

4 Discusión

A pesar de que las características de entrada son magnitudes relativas, aparecen diferencias significativas entre los diversos corpus (véase Tabla 2), que afectan tanto a μ como a σ . Estas diferencias se esperaban, independientemente del análisis multilingüe, porque las condiciones de grabación son diferentes, y éstas tienen un claro impacto en los valores de las magnitudes de las características de entrada. Así, por ejemplo, los valores de las propiedades relativas a F0 de *esma* se han recogido con una laringógrafo y los valores de F0 del corpus *Boston* se han obtenido mediante un algoritmo de extracción del pitch, lo que lleva a valores mucho menos estables.

El segundo punto de discusión que surge de la Tabla 2 es que, a la vez que las diferencias entre el corpus español y el inglés son claras, las diferencias entre los diversos sub-corpus en inglés también son importantes. La normalización de las características de entrada es por tanto una necesidad en este trabajo no sólo para reducir las diferencias que tienen su origen en las condiciones de grabación y procesamiento, sino también para hacer que la comparación entre varios idiomas sea posible. En (Escudero/González/Vivaracho/Cardeñoso/Aguilar 2011), se discute sobre el impacto en los resultados de las tasas de clasificación cuando la entrada está normalizada y cuando no lo está (más de diez puntos de precisión se pueden perder en el escenario multilingüe).

Tal y como se indica en la sección anterior, las tasas de clasificación son satisfactorias. Los escenarios multilingüe muestran menores tasas de identificación que los obtenidos en los escenarios monolingüe. Sin embargo, esta disminución es comparable a la obtenida en los escenarios en los que se cruzan locutores, a pesar de que los locutores pertenecen al mismo corpus y las condiciones de grabación son similares y el lenguaje el mismo.

La diferencia entre las tasas de clasificación inter-locutor tiene su origen en el papel diferente que las características prosódicas tienen de cara a la caracterización de los acentos. Este papel es dependiente de locutor como se muestra en la Tabla 4 y en la Figura 1, de manera que los diferentes locutores parecen utilizar de forma diferente las características prosódicas cuando realizan los acentos. Cuanto más parecido es el papel que juegan las características prosódicas entre dos locutores, mayor tasa de reconocimiento. Este hecho parece ser tan importante como el idioma en el que se producen las locuciones.

En (Escudero/González/Vivaracho/Cardeñoso/Aguilar 2011) se analizan las confusiones más comunes, es decir, situaciones en las que el clasificador comete un error al establecer la etiqueta a una palabra dada. Este análisis se llevó cabo mediante la comparación de las predicciones de los clasificadores con respecto a las etiquetas asignadas por un equipo de etiquetadores manuales ToBI (Prieto/Rosedano 2010). El resultado es interesante porque encontramos que tanto los clasificadores entrenados en condiciones multilingüe como los clasificadores entrenados en entornos monolingüe comparten la mayoría de los confusiones comunes. Así por ejemplo, el error más común es el de clasificar como acento el tono L *, lo

cual representa más del 35% de todos los desacuerdos. Además, los cuatro desacuerdos más comunes, representan más del 80% del total de los desacuerdos, y son compartidos por ambos clasificadores. Una vez más, los cuatro acuerdos más comunes que representan más del 80% de los acuerdos también son los mismos para ambos clasificadores. Este resultado evidencia un comportamiento similar de los clasificadores, y es un argumento a favor del uso del etiquetado de los eventos prosódicos en entornos multilingüe en combinación con una revisión posterior de los resultados por parte de etiquetadotes humanos.

5 Conclusiones y trabajos futuros

En este trabajo se ha presentado una experiencia de clasificación multilingüe en la que un clasificador es entrenado con datos de énfasis en una lengua para etiquetas las palabras acentuadas en otra lengua. Se ha realizado un análisis cuantitativo de los resultados pero, sobre todo, un análisis de los contrastes entre lenguas que pueden suponer dificultades en este proceso.

Las tasas relativas de identificación son altas, mientras que las confusiones son consistentes con las expectativas de acuerdo a la forma de los diferentes tonos ToBI del español. La introducción de técnicas de adaptación a locutor, más el uso de propiedades prosódicas representativas se manifiestan como una necesidad para dar validez a este proceso. El uso de características prosódicas más potente puede ser una opción para mejorar los resultados en trabajos futuros. En estos momentos estamos trabajando en la inclusión de otras características más expresivas, tales como parámetros de interpolación Bézier. Tilt y parámetros de Fujisaki para mejorar los resultados (González/Vivaracho/Escudero/Cardeñoso 2010). Además, también se está estudiando la inclusión de las técnicas de fusión de expertos para mejorar los resultados de clasificación, ya que las predicciones de los dos clasificadores pueden ser complementarias en algunos casos.

6 Agradecimientos

Este proyecto ha sido realizado en el marco de los proyectos de investigación VA322A11-2 financiado por la Consejería de Educación de la Junta de Castilla y León y por el Ministerio de Ciencia e Innovación, del Gobierno de España FFI2008-04982-C003-02 y FFI2011-29559-C02-01

Referencias Bibliográficas

Ananthakrishnan, Sankaranarayanan/ Narayanan, Shri. (2008): "Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence", en: *IEEE Transactions on Audio, Speech, and Language Processing,* 16(1), 216–228

Bonafonte, Antonio/Moreno, Asunción (2008): *Documentation of the upc-esma spanish database.* Tech. rep., TALP Research Center, Universitat Politecnica de Catalunya, Barcelona, Spain

Chawla, Nitesh Bowyer/Kevin, Hall Lawrence/ Kegelmeyer, Philip (2002) "Smote: Synthetic minority over-sampling technique", en : *Journal of Artificial Intelligence Research* 16, 321–357

Escudero, David/Cardeñoso, Valentín (2007): "Applying data mining techniques to corpus based prosodic modeling speech", en: *Speech Communication* 49, 218–229

Escudero, David/González, César/ Vivaracho, Carlos/ Cardeñoso, Valentín/ Aguilar, Lourdes (2011) "Analysis of inconsistencies in cross-lingual automatic tonal ToBI labelling", en: *Proceedings of TSD 2011 (International Conference on Text, Speech and Dialogue)* 41-48

González, César/Vivaracho, Carlos/Escudero, David/Cardeñoso, Valentín (2010): "On the Automatic ToBI Accent Type Identification from Data". en: *Proceedings of Interspeech 2010*

Hall, Mark/Frank, Eibe/Holmes, Geoffrey/Pfahringer, Bernhard/Reutemann, Peter/Witten, Ian (2009) "The WEKA Data Mining Software: An Update" en: *SIGKDD Explorations* 11, 10–18

Meteer, Marie/Schwartz, Richard/Weischedel, Ralph (1991) "Post: Using probabilities in language processing." en: *International Joint Conferences on Artificial Intelligence.* 960–965

Ostendorf, Mary/Price, Patti/Shattuck, Stephany (1995) *The Boston University Radio News Corpus*. Tech. rep., Boston University

Prieto, Pilar/Rosedano, Paolo (2010) *Transcription of Intonation of the Spanish Language*. LINCOM Studies in Phonetics 06

Rangarajan, Sridhar/Bangalore, Sridhar/Narayanan, Shrikanth (2008) "Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework", en *IEEE Transactions on Audio, Speech, and Language Processing* 16(4), 797–811

Syrdal, Ann/Hirschberg, Julia/McGory, Julie/Beckman, Mary (2001) "Automatic ToBI prediction and alignment to speed manual labeling of prosody". *Speech Communication* (33), 195–151

Vivaracho, Carlos/Simon, Arancha (2010) "Improving the performance for imbalanced data sets by means of the ntil technique", en: *IEEE International Joint Conference on Neural Networks* 18-23

Wightman, Colin Ostendorf, Mari (1994) "Automatic labeling of prosodic patterns" en: *IEEE Transactions on Speech and Audio Processing* 2(4), 469–481

Witten, Ian/Frank, Eibe (1999) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann