

# Introducción a las Tecnologías del Habla

## Trabajo Práctico 3: Aprendizaje Automático

Departamento de Computación,  
Facultad de Ciencias Exactas y Naturales,  
Universidad de Buenos Aires

Leandro Lovisolo  
LU 645/11

Segundo Cuatrimestre de 2012

### Introducción

El objetivo de este trabajo práctico es construir un sistema de reconocimiento automático del género de una persona a partir de una grabación corta de su habla, aplicando técnicas de aprendizaje automático.

Para la realización del sistema se dispone de un corpus de grabaciones recolectadas por todos los alumnos de esta materia durante el TP 1, de las cuales se extrajeron el género del hablante y un conjunto de atributos acústicos que serán utilizados como referencia para entrenar el sistema y evaluar su eficacia.

El sistema deberá implementarse sobre la suite de aprendizaje automático Weka<sup>1</sup>, en la que se deberá construir un clasificador que tome como entrada los atributos acústicos de una grabación, y decida en base a estos el género de la persona.

En primer lugar, se deberá implementar como *sistema baseline* un clasificador de reglas RIPPER<sup>2</sup> utilizando como único atributo la media de la frecuencia fundamental del hablante. En el TP1 habíamos visto que la diferencia de este atributo para cada género era significativa y grande. Ahora veremos cuál es su poder predictivo en esta tarea.

Finalmente, se deberá experimentar con diferentes clasificadores y diferentes conjuntos de atributos, en busca de una configuración que arroje buenos resultados. La tasa de aciertos deberá ser mayor o igual a 94%.

### Materiales y métodos

Las instancias en la base de datos corresponden a los segmentos del habla sin pausas (inter-pausal units o IPUs) de todas las grabaciones. Cada instancia registra el género del hablante y 1582 atributos acústicos.

Los atributos acústicos en la base de datos fueron extraídos de los archivos de audio con la herramienta openSMILE<sup>3</sup>, usando la configuración para el INTERSPEECH 2010 Paralinguistic Challenge<sup>4</sup>, y almacenados en formato ARFF para facilitar su lectura desde Weka. Para más información, ver las páginas 30 y 31 del openSMILE book<sup>5</sup>.

La base de datos de atributos acústicos, junto con el enunciado completo del TP, pueden descargarse desde la siguiente URL: <http://habla.dc.uba.ar/gravano/ith-2012/tp3/>

### Sistema baseline

Se empleó un clasificador `rules.JRip` con opciones de configuración por defecto y cross-validation de 10 folds.

El atributo utilizado fue `F0Final_sma_amean` (media de la frecuencia fundamental.)

---

<sup>1</sup><http://www.cs.waikato.ac.nz/ml/weka/>

<sup>2</sup>Repeated Incremental Pruning to Produce Error Reduction (RIPPER.) Ver <http://wiki.pentaho.com/display/DATAMINING/>

JRip

<sup>3</sup><http://opensmile.sourceforge.net/>

<sup>4</sup><http://emotion-research.net/sigs/speech-sig/paralinguistic-challenge>

<sup>5</sup>[http://sourceforge.net/projects/opensmile/files/opensmile\\_book\\_1.0.0.pdf](http://sourceforge.net/projects/opensmile/files/opensmile_book_1.0.0.pdf)

## Resultados

Se obtuvo un porcentaje de instancias correctamente clasificadas del **86.9315%**, con la siguiente matriz de confusión:

=== Confusion Matrix ===

```

  a   b   <-- classified as
660  94 |   a = f
110 697 |   b = m

```

## Experimentos realizados

Se evaluaron los clasificadores `rules.JRip`, `trees.J48`, `bayes.NaiveBayes` y `functions.SMO` con cross-validation de 10 folds, utilizando tres subconjuntos diferentes de atributos acústicos de la base de datos:

- La base de datos completa.
- El subconjunto hallado utilizando el método de búsqueda **GreedyStepwise**, con evaluador de atributos `ClassifierSubsetEval` y clasificador `functions.SMO`:

**mfcc\_sma[0].linregc1** coeficiente cepstral 0 en las frecuencias de Mel: pendiente de la aproximación lineal.

**mfcc\_sma[0].linregerrQ** coeficiente cepstral 0 en las frecuencias de Mel: error cuadrático de la aproximación lineal.

**mfcc\_sma[4].skewness** coeficiente cepstral 4 en las frecuencias de Mel suavizado<sup>6</sup>: asimetría estadística.

**mfcc\_sma[6].linregc2** coeficiente cepstral 6 en las frecuencias de Mel suavizado: punto de corte con el eje *y* de la aproximación lineal.

**mfcc\_sma[10].quartile1** coeficiente cepstral 10 en las frecuencias de Mel suavizado: primer cuartil.

**mfcc\_sma[12].amean** coeficiente cepstral 12 en las frecuencias de Mel suavizado: media aritmética.

**mfcc\_sma[14].quartile2** coeficiente cepstral 14 en las frecuencias de Mel suavizado: segundo cuartil.

**lspFreq\_sma[1].linregerrA** par 1 de líneas espectrales de frecuencias computados a partir de coeficientes LPC (linear predictive coding) suavizado: error lineal de la aproximación lineal.

**lspFreq\_sma[7].quartile3** par 7 de líneas espectrales de frecuencias computados a partir de coeficientes LPC (linear predictive coding) suavizado: tercer cuartil.

**F0finEnv\_sma\_amean** envolvente de la frecuencia fundamental suavizada: media aritmética.

**mfcc\_sma\_de[9].percentile99.0** diferencial del coeficiente cepstral 9 en las frecuencias de Mel suavizado: 99-percentil.

**lspFreq\_sma\_de[2].linregerrQ** diferencial del par 2 de líneas espectrales de frecuencias computados a partir de coeficientes LPC (linear predictive coding) suavizado: error cuadrático de la aproximación lineal.

**lspFreq\_sma\_de[6].quartile2** diferencial del par 6 de líneas espectrales de frecuencias computados a partir de coeficientes LPC (linear predictive coding) suavizado: segundo cuartil.

**F0final\_sma\_upleveltime75** envolvente de la frecuencia fundamental suavizada: porcentaje de tiempo que la señal está por encima del 75%.

- El subconjunto hallado utilizando el mismo método y evaluador anteriores, pero reemplazando el clasificador por `rules.JRip`:

**mfcc\_sma[10].quartile1** coeficiente cepstral 10 en las frecuencias de Mel suavizado: primer cuartil.

**mfcc\_sma[12].quartile1** coeficiente cepstral 12 en las frecuencias de Mel suavizado: primer cuartil.

**logMelFreqBand\_sma[0].quartile3** potencia logarítmica de la banda 0 en las frecuencias de Mel suavizada: tercer cuartil.

**lspFreq\_sma[4].linregerrA** par 4 de líneas espectrales de frecuencias computados a partir de coeficientes LPC (linear predictive coding) suavizado: error lineal de la aproximación lineal.

**F0finEnv\_sma\_amean** envolvente de la frecuencia fundamental suavizada: media aritmética.

**shimmerLocal\_sma\_linregc1** desviaciones locales de amplitud entre períodos de pitch suavizadas: pendiente de la aproximación lineal.

<sup>6</sup>Suavizado por un filtro promedio móvil con ventana de longitud 3.

## Resultados

A continuación se presentan los porcentajes de aciertos obtenidos con todas las combinaciones de clasificadores y subconjuntos de atributos acústicos.

Clasificador	Todos los atributos	Búsqueda GreedyStepwise con ClassifierSubsetEval	
		Clasificador functions.SMO	Clasificador rules.JRip
rules.JRip	96.2844%	94.5548%	95.3235%
trees.J48	93.5959%	94.6188%	95.0673%
bayes.NaiveBayes	86.6752%	94.6188%	85.3299%
functions.SMO	99.0391%	97.7578%	91.672%

## Mejor sistema desarrollado

La mejor configuración hallada fue un clasificador `functions.SMO` con cross-validation de 10 folds, que toma todos los atributos acústicos de la base de datos.

## Resultados

Se clasificaron correctamente el **99.0391%** de las instancias, y se obtuvo la siguiente matriz de confusión:

=== Confusion Matrix ===

```

  a    b    <-- classified as
747    7 |    a = f
  8 799 |    b = m
```