

# Extracción de embocadura en Aliento/Arrugas

PROYECTO FINAL - PROCESAMIENTO DIGITAL DE SEÑALES DE AUDIO - CURSO 2016  
MAESTRÍA EN INGENIERÍA ELÉCTRICA del *Instituto de Ingeniería Eléctrica, Facultad de Ingeniería, Universidad de la República.*

Juan Braga

12 de diciembre de 2016

## Resumen

El presente trabajo tiene como objetivo principal la evaluación de características de diversa naturaleza, en la extracción automática de embocadura a partir de muestras de audio. Se trabaja sobre Aliento/Arrugas, obra para flauta con técnicas extendidas, del compositor contemporáneo Marcelo Toledo.

## 1. Introducción

La Flauta travesa, perteneciente a la familia de *Instrumentos de Madera*, es la única que no utiliza *lengüeta* o *caña* en su embocadura. Su mecanismo de producción de la exitación periódica difiere del resto, ya que esta basado en la turbulencia generada por la colisión del flujo del aire con el bisel de su agujero. La exitación periódica pone a resonar a la columna de aire, y la altura es definida mediante la posición de las llaves del instrumento, característica que comparte con el resto de la familia de *Maderas*.

Además de la ejecución de alturas como material sonoro, los compositores académicos contemporáneos, exploran las capacidades sónicas del instrumento. Actualmente existen un montón de técnicas modernas, claramente definidas y reproducibles a las que se denominan extendidas. Diversos recursos técnicos forman parte del repertorio actual de la flauta y los intérpretes deben expandir sus habilidades con el instrumento.

La extracción de contenido musical en este tipo de obras no se abarca solamente con estimación de alturas y duraciones, a diferencia del lenguaje tradicional de la flauta. El repertorio contemporáneo define un área del *MIR* (por su sigla en inglés *Music Information Retrieval*) desafiante y de espectro más amplio. Además es un área de estudio donde no existe mucho camino andando, habiendo tierra fértil para explorar y lograr contribuciones.

El cometido del presente trabajo es el acercamiento a este tipo de problemas, con objetivo la extracción automática de embocadura a partir de grabaciones de la obra contemporánea Aliento/Arrugas. En lo que sigue, Sección 1.1, se define con más detalle el término embocadura, y se hace un análisis de sus principales implicancias en el sonido de la flauta. Además en la Sección 1.2 se describe en mayor profundidad la pieza musical Aliento/Arrugas. Luego en la Sección 2 se define el problema y la estrategia de resolución, se detalla la base de datos y se da un breve marco teórico de las características a utilizar. La Sección 3 está dedicada a comentar los resultados y algunas consideraciones, para finalmente concluir el trabajo en el Sección 4. Por otro lado en el Anexo se puede ver una representación gráfica de las predicciones para comparación visual con el ground truth.

### 1.1. Embocadura

El término embocadura refiere al aparato de producción de la exitación de la columna de aire, en conjunto a la técnica de soprido (Piston, 1955, Capítulo 6). Por ejemplo en el caso particular de la flauta ejecutada con técnica tradicional, los labios dirigen el flujo de aire directamente al bisel en el hueco del instrumento. De esta forma la turbulencia producida por la colisión, genera una exitación periódica en la columna de aire que provoca la resonancia del instrumento y un sonido tonal.

La embocadura es un elemento determinante del material sonoro ejecutado, siendo perceptible de forma auditiva a través de variaciones en la dinámica, altura y timbre (Dick, 1975, Capítulo 2). Las características sonoras quedan determinadas por los siguientes parámetros físicos de la ejecución del instrumento:

- **Ángulo de la flauta:** Por un lado afecta la altura de la ejecución. Al girar hacia el intérprete la altura baja, por el contrario sube al girar en el otro sentido. Por otro lado genera cambios en el timbre del material sonoro. Al girar hacia afuera (sentido opuesto al intérprete) mas allá del ángulo normal de ejecución, el sonido se vuelve primero más brillante y luego aumenta la prominencia del componente de ruido, en inglés

se lo define como *Breathy*, se lo puede traducir al español como *Respirado*. Sin embargo al girar hacia el intérprete aumenta la energía de los parciales altos y disminuye la fundamental generando un sonido que se puede definir metafóricamente como *Filoso* o *Edgy* por su denominación en inglés.

- **Apertura de los labios:** La apertura de los labios determina la dispersión del flujo de aire. Aperturas pequeñas producen flujos puntualas, disminuyendo la dinámica y clarificando el sonido. Del otro lado aperturas mayores aumentan la intensidad y la naturaleza ruidosa.
- **Posición de los labios:** Una posición correcta de los labios genera que la embocadura tenga gran control del sonido. Si bien la posición de los labios, y los movimientos de los mismos en la ejecución es un aspecto personal del ejecutante, existen dos tipos básicos. Alturas bajas y/o dinámicas intensas aumentan con el movimiento de los bordes de los labios hacia afuera generando casi una sonrisa en el intérprete. En la segunda posición de los labios, los bordes se mueven hacia abajo en vez de hacia afuera, teniendo un efecto similar al mencionado anteriormente.
- **Presión de aire:** La presión de aire es controlada por el diafragma. Determina el nivel dinámico de la ejecución. La intensidad del aire es proporcional a la intensidad de la ejecución. Además afecta la altura del material sonoro, presiones de aire altas tienden a elevar la nota, mientras que presiones menores la disminuyen.

## 1.2. Aliento/Arrugas de Marcelo Toledo

Aliento/Arrugas es una obra para flauta traversa solista, compuesta por el argentino Marcelo Toledo. Incluye una cantidad de sonoridades exóticas mediante la ejecución del instrumento a través de técnicas extendidas. Según el compositor la intención detrás es la exploración sonora del instrumento utilizando la respiración del intérprete como elemento de expresión orgánica (Candelaria et al., 2005).

El compositor utiliza como recurso expresivo tres tipos de embocadura para ejecución del instrumento. Se diferencian por cambios en la posición de los labios y el ángulo de la flauta (ver 1.1), en otras palabras el ángulo entre el flujo de aire frente al bisel de la embocadura. Se enlista a continuación los nombres, manteniendo su denominación en Inglés (idioma utilizado en la partitura de la obra). Además en la Figura 1 se observa la notación utilizada por el compositor en la partitura de Aliento/Arrugas.

- *Normal Embouchure*: Embocadura clásica de la flauta, donde el flujo de aire frente al bisel de la embocadura genera la exitación con pulsos periódicos de la columna de aire.
- *Blow Hole Covert*: El flujo de aire ingresa directo al tubo de la flauta, sin generar turbulencia contra el bisel de la embocadura. Los labios cubren el agüjero del instrumento.
- *Breathy Embouchure*: La flauta se encuentra rotada hacia el lado contrario del intérprete, tomando como referencia la embocadura normal. Genera sonidos con orientación tonal pero con un gran componente ruidoso.

Figura 1: Notación de las embocaduras se observa en la parte superior de los sistemas. (a) *Blow Hole Covert*. (b) *Breathy Embouchure*. (c) *Normal Embouchure*. Fragmentos extraídos de la partitura de Aliento/Arrugas.

## 2. Definición del Problema

Teniendo en cuenta que la embocadura es un elemento determinante del material sonoro ejecutado perceptible de forma auditiva (Sección 1.1), se propone la extracción automática del tipo de embocadura a través del análisis computacional de grabaciones de la obra.

### 2.1. Estrategia de resolución

Se propone la resolución del problema con un enfoque de reconocimiento de patrones. Se procesa el audio como un *Bag of Frames* a partir del computo de descriptores numéricos. El principal desafío y cometido del presente trabajo es encontrar los descriptores que extraigan las diferencias en la naturaleza sonora y permitan la separación de las embocaduras en el espacio de características.

### 2.2. Conjunto de Datos

Se cuenta con 5 grabaciones de diferentes intérpretes de la obra Aliento/Arrugas. Los intérpretes son: Pablo Somma, Emma Resmini, Claire Chase, Juan Pablo Quinteros y Ulla Suokko. Los archivos de audio se etiquetaron utilizando el software *Sonic Visualiser* (Cannam et al., 2010) dividiendo los archivos de audio en 5 clases:

- Silencio.
- Silencio con respiración del intérprete.
- Sonido generado con *Blow Hole Covert*.
- Sonido generado con *Breathy Embouchure*.
- Sonido generado con *Normal Embouchure*.

Las grabaciones de Claire Chase y Juan Pablo Quinteros que se obtuvieron para el presente trabajo sufrieron un proceso de compresión con pérdida, por lo que estos datos reciben un tratamiento distinto. No se utilizan para entrenar los algoritmos de clasificación, solo se utilizan como datos de test. Por lo que folds son de la siguiente forma:

- Cuando la grabación de test es la de Ulla Suokko, Pablo Somma o Emma Resmini, se entrena con las otras dos restantes. Metodología de *Leave One Out* por su denominación en Inglés.
- Por otro lado cuando la grabación de test es de Claire Chase o Juan Pablo Quinteros, el conjunto de entrenamiento esta compuesto por las tres grabaciones sin pérdida (i.e. la de Ulla Suokko, Pablo Somma y Emma Resmini).

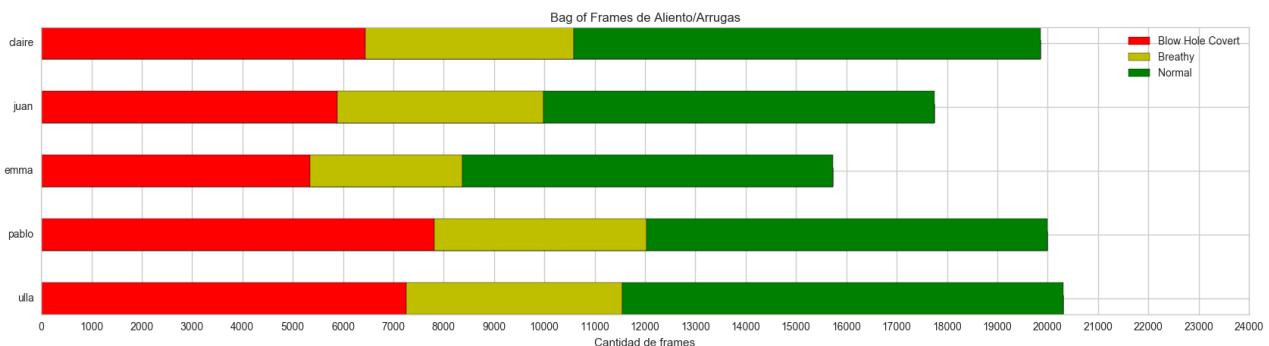


Figura 2: Detalle de la composición del *bag of frames* de embocaduras.

En la Figura 2 se observa en detalle las cantidades de frames de embocaduras en la base de datos. Se puede observar según el intérprete la proporción de las clases. Se observa que la clase mayoritaria es la *Normal embouchure* de numero comparable *Blow Hole Covert* y sensiblemente menor es la dada por la embocadura *Breathy*.

En lo que sigue se utilizan únicamente las clases asociadas a cada una de las embocaduras. Queda por fuera del alcance de este trabajo, una etapa de pre-procesamiento para la segmentación del audio en fragmentos de actividad de la flauta y silencios (este problema se conocido como *Activity Detection* por su denominación en Inglés).

## 2.3. Extracción de características

Se enlistan a continuación las características que se evalúan en la extracción automática de embocadura. Además se describe brevemente sus principales atributos.

### 2.3.1. Mel-Frequency Cepstral Coefficients (MFCC)

Los *Coeficientes Cepstrales de Frecuencia-Mel* fueron introducidos por Davis and Mermelstein (1980) en la resolución del problema de reconocimiento del hablante a partir de señales de voz (*Speaker Recognition* su denominación en Inglés). Estos coeficientes como características de un sistema de reconocimiento automático del hablante han demostrado tener de los mejores desempeños (Quatieri, 2002, Capítulo 14). A partir de ahí han sido utilizados en diversas problemáticas de clasificación que no involucran señales de voz hablada, con buenos resultados también como es el caso de reconocimiento de instrumentos (Klapuri and Davy, 2007, Capítulo 6). Su fortaleza radica en la incorporación del modelado psicoacústico de la audición humana mediante un banco de filtros basados en la escala Mel (Stevens et al., 1937) y la decorrelación que presentan los datos en el dominio de las *quefrecencys*, dado por la aplicación de la Transformada Coseno. Son un buen descriptor para la extracción de aspectos tímbricos de la señal.

El cómputo de estas características cuenta con las etapas que se enlistan a continuación de manera conceptual:

- a. División de la señal en fragmentos mediante enventanado.
- b. Cálculo de la magnitud de la Transformada discreta de Fourier de tiempo corto (STFT).
- c. Fitrado de la señal con banco de filtros Mel.
- d. Cálculo de la energía para cada filtro del banco.
- e. Logaritmo de las energías.
- f. Transformada Coseno de los valores a la salida del Logaritmo.
- g. *Liftrado* de la señal resultante en el dominio de las *quefrecencys*, luego de la Transformada Coseno. Determina la cantidad de coeficientes, o en otras palabras la dimensión del espacio de caracterísicas.

El cálculo de los coeficientes MFCC tiene los siguientes parámetros determinantes de su desempeño: En primer lugar el largo de las ventanas, que define el compromiso entre resolución temporal ypectral. En segundo lugar la cantidad de filtros del banco de filtros Mel, que se puede pensar como un submuestreo de la resolución espectral ya determinada por el largo del enventanado. Por último el *liftrado* de la señal a la salida de la Transformada Coseno que determina la cantidad de coeficientes efectivos previos al clasificador.

### 2.3.2. Linear Prediction Coefficients (LPC)

La técnica de análisis de señales de tiempo discreto por predicción lineal, tiene su aplicación en diversas áreas del conocimiento. Es parte de un problema más general denominado *identificación de sistemas* desarrollado en el área de control para el análisis de sistemas dinámicos. Supone que la señal de análisis es la salida  $s[n]$  de un sistema lineal con entrada  $u[n]$ . Su fortaleza y versatilidad radica en la estimación de parámetros del sistema lineal que define el problema.

Un enunciado más general modela la señal de análisis como un proceso *Auto-Regresivo de Media Móvil (ARMA)* (Makhoul, 1975). En otras palabras, supone que la muestra actual de la señal de análisis puede ser expresada como una combinación lineal de las muestras pasadas de la salida, y la muestra actual y pasadas de la entrada:

$$s[n] = - \sum_{k=1}^p a_k s[n-k] + G \sum_{l=0}^p b_l u[n-l] \quad (1)$$

Ha sido utilizado para la resolución de problemas con señales de audio, en particular existe mucha literatura al respecto con la voz humana. Para voz hablada es de los métodos más poderosos, con diversas aplicaciones. La importancia de este método se basa tanto en la precisión de la estimación de parámetros del modelo de mecanismo de producción de voz, como en su relativo bajo costo computacional (Rabiner and Schafer, 1978, Capítulos 3 y 9).

Alineado con la utilización de LPC en problemas de señales de voz, se supone que es suficiente un modelo todo-polos para la extracción de características en el presente trabajo. De forma matemática a partir de la Ecuación 1 se escribe como  $b_l = 0$  con  $l = 1 \dots p$ . Los parámetros relevantes en el computo del descriptor son entonces, en primer lugar  $p$  asociado a la cantidad de polos del modelo AR y por otro lado el largo de las ventanas de análisis. Componiendo el vector de características por los coeficientes  $a_k$  con  $k = 1 \dots p$  (ver Ecuación 1).

### 2.3.3. Conjunto de características Espectrales y Armónicas

Se genera un vector compuesto por 5 características acústicas uni-dimensionales, para evaluación del desempeño en la extracción de embocadura. Entre los descriptores se optaron por 4 medidas espetrales y una medida de armónica de la señal de análisis, según la taxonomía de *features* acústicos propuesta en el libro de Klapuri and Davy (2007). Las características son:

- Voicing: Es una medida de periodicidad de la señal. Es el *feature* armónico del conjunto. Generalmente el Voicing se encuentra embebido en los algoritmos de extracción de pitch. En particular para el presente trabajo se computa como en la referencia: De Cheveigné and Kawahara (2002).
- Zero-Crossing Rate: Mide la cantidad de cruces por cero de la señal. Si bien es calculado en el dominio del tiempo, es una medida del contenido de alta frecuencia.
- Roll-off: Es el valor de frecuencia para el que la energía espectral acumulada supera una fracción denominada  $\lambda$ . En general  $\lambda$  se elige 95 % o 85 %.
- Centroid: Es el promedio en los bins de frecuencia ponderado por los valores de magnitud del espectro. Se puede pensar como el centro de masa en el espectro.
- Bandwidth: Es una medida de la dispersión spectral con respecto al centroide.

En todas las medidas acústicas recién mencionadas es de relevancia la elección del largo de la ventana análisis, que define el compromiso entre estacionariedad de la señal y resolución en frecuencia.

### 2.3.4. Octave-based Spectral Contrast (SC)

El *Contraste Espectral por Octavas* fue desarrollado en el trabajo publicado por Jiang et al. (2002). Tiene como cometido ser una medida de las características relativas del espectro de la señal de análisis. Extrae la diferencia entre la prominencia de los picos en el espectro y los valles en cada octava de análisis por separado. Ha tenido buenos resultados en el problema de clasificación de estilo musical.

El cómputo de estas características tiene las siguientes etapas, que se enuncian de forma conceptual:

- a. División de la señal en fragmentos mediante enventanado.
- b. Cálculo de la magnitud de la Transformada discreta de Fourier de tiempo corto (STFT).
- c. Fitrado de la señal con banco de filtros por octava.
- d. Cálculo de la diferencia entre la energía en un entorno de los picos y de los valles en cada una de las octavas.
- e. Logaritmo de las diferencias del paso anterior.
- f. Transformada *Karhunen-Loeve* para representación de las características en base ortonormal y decorrelación entre las dimensiones.

A diferencia de *MFCC* y *LPC* que realizan un promediado de la información espectral, estos descriptores extraen la información relativa, mediante la comparación de picos y valles por octava. Los parámetros relevantes son en primer lugar el largo de la ventana de análisis, el entorno de los picos y valles denominado  $\alpha$  en la literatura, y por último el número de octavas.

## 3. Experimentos

Se evalúa la capacidad de los descriptores presentados en la Sección 2.3 en la separación de embocaduras. Para esto se utilizan tres clasificadores distintos para minimizar el bías que pueda existir entre los datos y un algoritmo en particular. Se trabaja con los algoritmos: *Random Forest* (*trees=10*), *Support Vector Machine* (*kernel lineal*) y *K-Nearest Neighbors* ( $k=10$ ). En todos los casos se utilizan los parámetros por defecto ya que no es objetivo de este trabajo encontrar los valores óptimos de clasificación. La implementación se realiza mediante el módulo de *Python* llamado *Scikit Learn* (Pedregosa et al., 2011). En todos los casos los datos son preprocesados de manera de centrar en cero y escalar la varianza a uno, previo al clasificador.

Todos los experimentos se realizan con *5-fold cross validation* donde los folds son las diferentes interpretaciones de la pieza musical, como se detalla en la Sección 2.2. De esta forma se asegura que frames provenientes de la misma grabación no sean usados para train y test en un mismo experimento. Además los *features*: *MFCC*, *SC*, *Roll-off*, *Centroid*, *ZCR* y *Bandwith* se calculan utilizando el módulo de *Python* llamado *Librosa* (McFee et al., 2015).

### 3.1. Primer experimento: Mejor descriptor para extracción de embocadura

El propósito es cuantificar el poder de separación de las características y determinar cual tiene el mejor desempeño. Para tener una noción general del comportamiento de los descriptores, en todos los casos se utiliza más de una combinación de parámetros. Se eligieron de forma que sea suficiente para descartar los de menor desempeño. A continuación se enlistan los parámetros utilizados en cada caso.

- Características Espectrales y Armónicas: Se varían los largos de ventana y saltos de la siguiente forma (se detallan respectivamente): (a)  $11ms$  y 50 % de salto (256-128 muestras), (b)  $23ms$  y 50 % (1024-512 muestras) y por último (c)  $46ms$  y 50 % (2048-1024 muestras). En todos los casos anteriores se utilizó:
  - Voicing: Número de retardos: 250 muestras.
  - Roll-off:  $\lambda = 85\%$
  - Centroid, Bandwith y Zero-Crossing Rate: quedan definidos por el largo de la ventana de análisis y el salto.
- MFCC: Se computan con ventana de análisis de  $23ms$  y salto del 50 %, 40 bandas Mel y se liftra la señal para obtener: (a) 20 coeficientes, (b) 30 coeficientes y (c) 40 coeficientes.
- LPC: Se computan con ventana de análisis de  $23ms$  y salto del 50 % y numero de polos: (a) 10, (b) 20 y (c) 40.
- SC: Se computan con ventana de análisis de  $23ms$  y salto del 50 % y numero de bandas: (a) 3 y (b) 6.

### 3.2. Resultados

En lo que sigue se muestra el resultado del desempeño de los descriptores detallados en la Sección 2.3 para la extracción del tipo de embocadura.

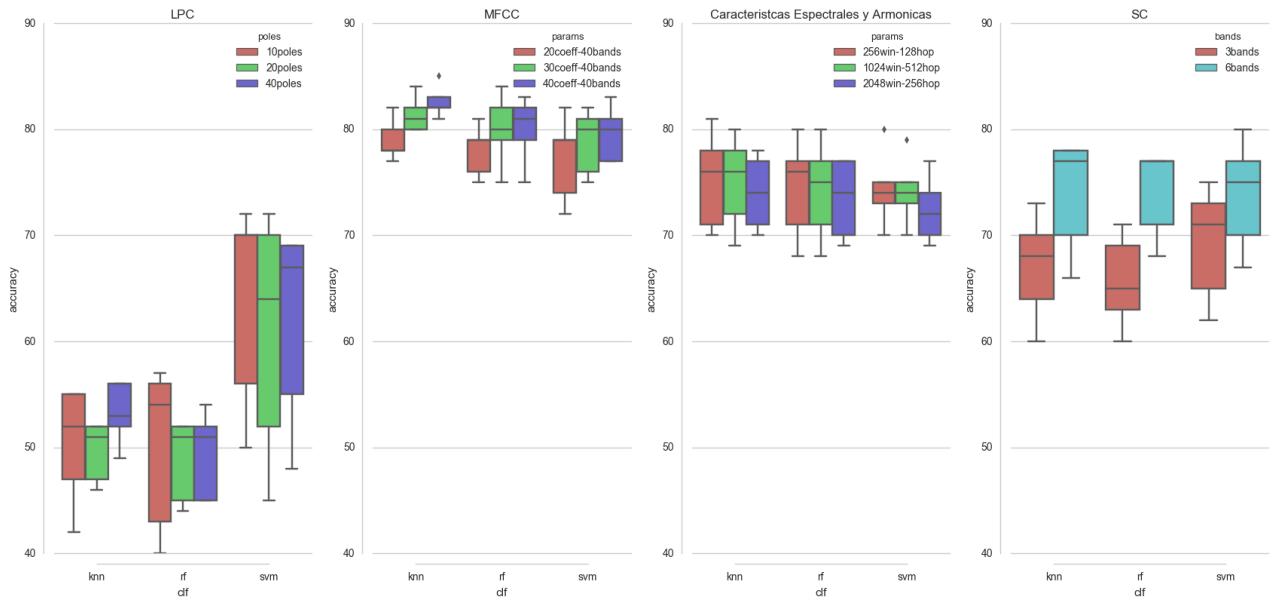


Figura 3: Boxplot para el accuracy de los algoritmos de clasificación. Se muestra los resultados de forma independiente por algoritmo de clasificación y según los parámetros de las características.

Se observa en la Figura 3 el comportamiento de las características en la separación de las clases del problema. Como medida de desempeño se utiliza la razón entre los frames bien clasificados y el total, denominada en Inglés como *accuracy*. Además en el eje horizontal se detalla el algoritmo de clasificación utilizado y en colores las distintas combinaciones de parámetros en la extracción de características.

En la Figura 3 se puede observar claramente que *LPC* es el que presenta peores resultados. Además de forma cualitativa se puede decir que la variación del número de polos del modelo, no afecta considerablemente el desempeño. Como resultado anecdótico a los fines del presente trabajo, se tiene que existe una mejora sustancial en el rendimiento de estos *features* con *SVM*.

Por otra parte el *feature SC* es de desempeño intermedio junto con *Características Espectrales y Armónicas* en el experimento. Para el caso de *SC* existe una mejora notoria al variar los parámetros. En otras palabras la comparación entre picos y valles tiene un poder descriptivo mayor de el problema, al dividir el espectro en 6 bandas con respecto a 3 bandas.

Del otro lado vemos que la variación de la ventana de análisis en *Características Espectrales y Armónicas* no generan un mejor rendimiento. Vale decir que su desempeño similar al de *SC* pesar de tener dimensiones correlacionadas como es el caso de *ZCR* y *Centroid*. Por lo que alguna estrategia de decorrelación previo al clasificador podría mejorar el desempeño de este conjunto de medidas.

También se observa que *MFCC* es el *feature* de mejor desempeño para la resolución del problema. Como contra partida, frente a *SC* y *Características Espectrales y Armónicas* la dimensión del espacio de características es mayor, resultando en un costo computacional superior.

Otro análisis relevante del experimento esta dado por las matrices de confusión. Para detallar los resultados se deja de lado *LPC* de pobre rendimiento, y se computan las matrices de confusión con la predicción realizada con *KNN* a la grabación de Emma Resmini como conjunto de test. En la Figura 4 se observa las matrices de confusión respectivas, todos los valores son en porcentaje, relativos a la cantidad de frames en la clase.

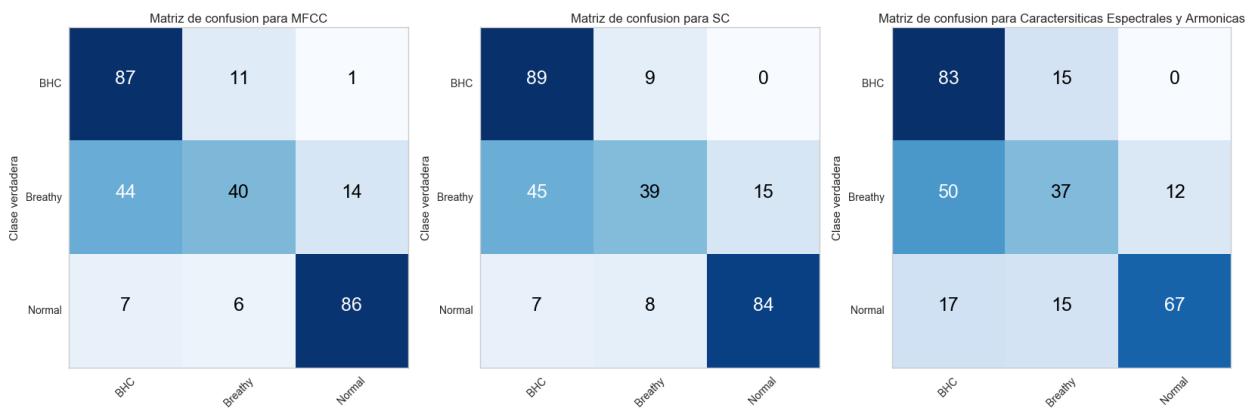


Figura 4: Matrices de confusión para las características *MFCC*, *SC*, y *Características Espectrales y Armónicas* de izquierda a derecha respectivamente. Para todos los casos el algoritmo de clasificación es *KNN*.

Los resultados demuestran, independientemente de los *features*, que las clases *Blow Hole Covert* y *Normal Embouchure* se separan frente al resto. No es el caso de *Breathy* que principalmente se confunde con *Blow Hole Covert*. Es razonable ya que se puede pensar como el caso intermedio desde el punto de vista acústico, entre las tres clases. Queda planteado entonces el punto débil en la extracción de embocadura de las características propuestas. De ahora en más se trabaja con el *feature MFCC* de mejor desempeño.

### 3.3. Segundo experimento: Blow Hole Covert Vs. Breathy

En lo que sigue se evalúa nuevamente el desempeño de *MFCC* en la separación de las clases pero con una versión reducida del problema, teniendo en cuenta solamente las dos clases problemáticas: *Blow Hole Covered* y *Breathy Embouchure*.

Por simpleza se trabaja solamente con *MFCC* computado con 40 bandas Mel y 20 coeficientes. En la Figura 5 se observa el *accuracy* para los distintos algoritmos de clasificación.

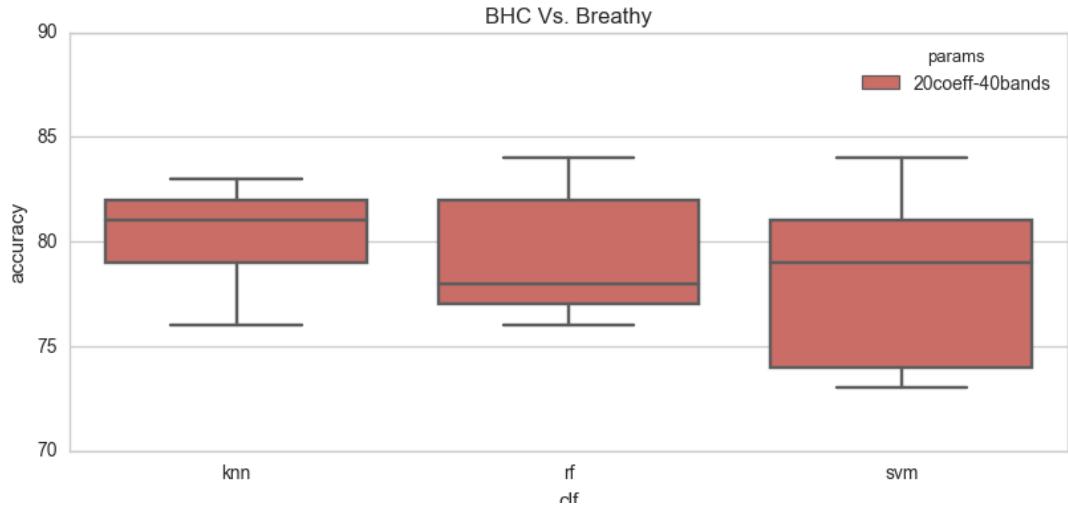


Figura 5: Matrices de confusión para las clases *BHC* Vs. *Breathy* generadas con características MFCC y el clasificador KNN.

El desempeño es similar al problema completo de tres clases y no existen grandes diferencias en el rendimiento según el algoritmo de clasificación. En la Figura 6 se observa la matriz de confusión para la predicción realizada con KNN a la grabación de Emma Resmini como conjunto de test, además se normaliza los resultados del experimento de la Sección 3.1 para realizar la comparación.

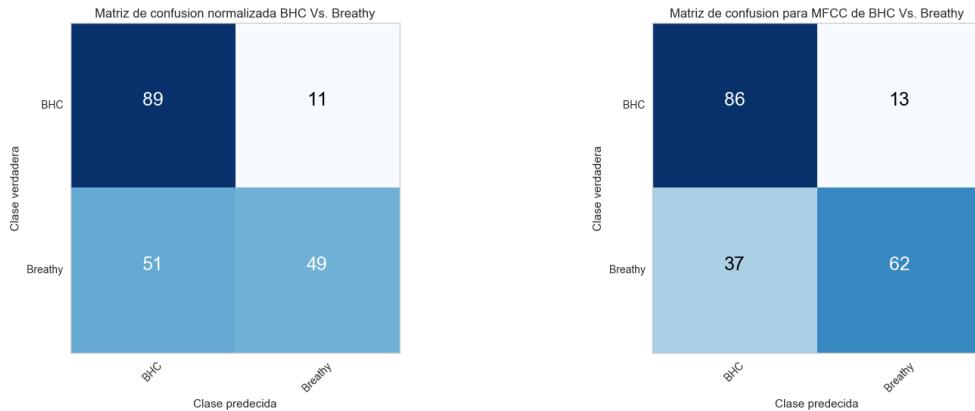


Figura 6: Matrices de confusión para las clases *BHC* Vs. *Breathy* generadas con características MFCC y el clasificador KNN.

Si bien sigue siendo considerable la confusión con un 37 % de elementos de *Breathy* clasificados como *Blow Hole Covert*, hay una mejora en comparación al experimento anterior, acertando ahora en la mayoría de los casos.

Una estrategia de dos etapas de clasificación en cascada mejoraría los resultados con respecto al Experimento 3.1 a cambio de mayor costo computacional.

### 3.4. Refinamiento de la extracción de características basada en MFCC

En lo que sigue se buscan los parámetros de *MFCC* que logran el resultado óptimo para el problema dado. Recordando la matriz de confusión de la Figura 4, para el mejor caso (*MFCC + KNN*) existe un 13 % de frames ejecutados con *Normal Embouchure* que fueron mal clasificados.

Mientras *MFCC* es una buena medida del aspecto tímbrico del material sonoro por extraer la envolvente espectral, a priori no contiene información de la periodicidad de la señal de análisis. Por lo que se propone agregar el cálculo de *Voicing* como una dimensión más del vector de características y evaluar si existe disminución en la confusión de la clase *Normal Embouchure*.

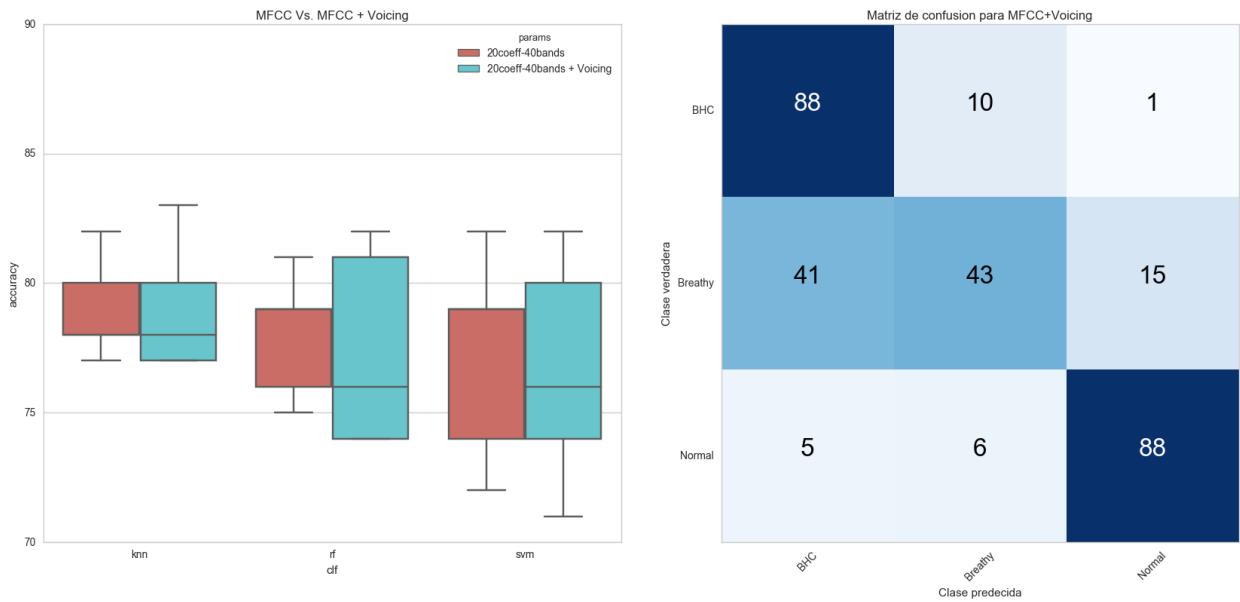


Figura 7: *Accuracy* y matriz de confusión para la evaluación comparativa del agregado de *Voicing* a las características *MFCC*

Se observa en la Figura 7 que tanto el *accuracy*, como la confusión de la clase *Normal Embouchure*, no denotan cambios relevantes de forma cualitativamente.

En lo que sigue se evalúa el desempeño de *MFCC* variando el largo de la ventana de análisis en los valores *11ms*, *23ms* y *46ms*. En todos los casos el salto entre ventanas es del 50 %.

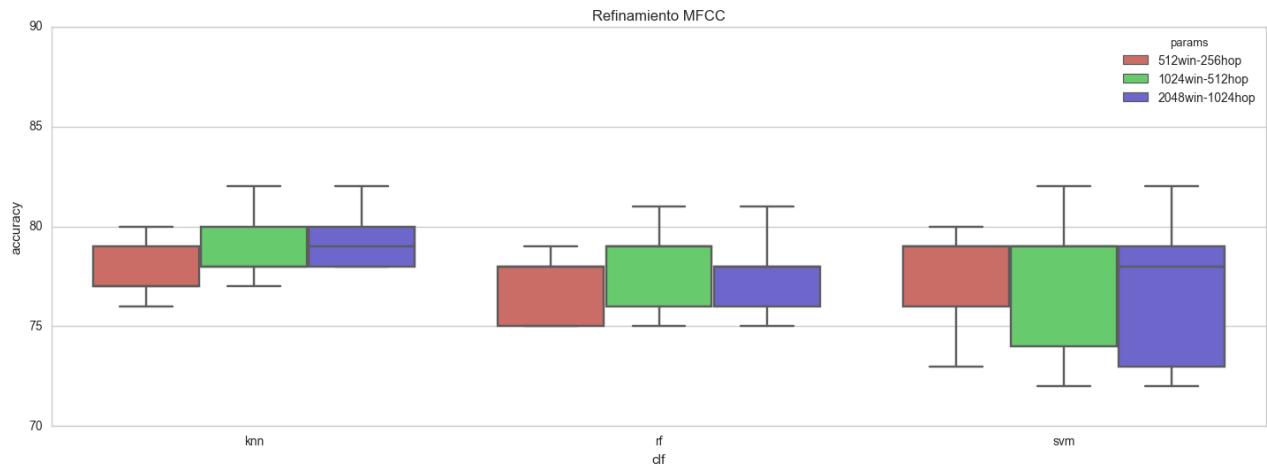


Figura 8: Accuracy de *MFCC* al variar el largo de ventana de análisis.

Como se observa en la Figura 8 no existen cambios relevantes en el desempeño al variar el largo de la ventana. Se puede decir que para estos valores la variación tanto en la resolución temporal como en la espectral no es suficiente como para denotar diferencias en el desempeño.

Por ultimo decir que la mejor combinación de parámetros para el cómputo de *MFCC* esta dada por una ventana de *23ms*, 40 bandas Mel y sin lifrado en el dominio de las quefrecyys, resultando en un vector 40 dimensiones, siendo alta con respecto a las que se han manejado en el presente trabajo (ver Figura 3). Vale restaltar que usualmente se liftra la señal ya que la información tímbrica no se encuentra en las altas quefrecyys pero los resultados para estos datos sugieren lo contrario.

## 4. Conclusiones

Las características evaluadas en el presente trabajo no fueron adecuadas para la separación de las clases *Blow Hole Covert* y *Breathy*. Con un estudio mas minucioso sobre la naturaleza acústica de estos dos embocaduras se podría diseñar algún *feature* que desambigüe la decisión.

Si bien se concluye que se debe mejorar la separación entre las embocaduras *Blow Hole Covert* y *Breathy*, existen pasajes de transición, inicios y finales de frases musicales, en los que se vuelve ambigua su naturaleza acústica, y se deben tener en cuenta al momento del etiquetado de las embocaduras.

Desde otro punto de vista la ambigüedad entre las clases sugiere que el enfoque basado en el aspecto tímbrico estimado como la envolvente espectral, no es suficiente. Si bien la embocadura varia el material sonoro y, en un sentido amplio, la composición tímbrica, hay que tener en cuenta que el instrumento físico no cambia. Por lo que el resonador es estacionario y mínima la variación de la estimación de la envolvente espectral a lo largo el tiempo. Esto explicaría por un lado porque los *features* *LPC* tuvieron un pobre desempeño y por otro porque el accuracy óptimo se logra con *MFCC's* sin liftrar. Queda planteada la hipótesis de que la información relevante está en la exitación generada por el intérprete y no por las características del resonador dadas por la envolvente espectral.

Por último vale mencionar que la estrategia de *bag of frames* es exigente, ya que se descarta la información temporal de la señal de audio, dejando de lado toda la información a priori. Es mucha la información relevante dada por ser grabaciones de audio de una interpretación musical, de un estilo definido y por si fuera poco con partitura disponible.

### 4.1. Trabajo a futuro

- Enfocar la resolución del problema desde la estimación de la exitación generada por el intérprete, y evaluar el desempeño con esta estrategia.
- Segmentación automática del audio en fragmentos de actividad de la flauta y silencios, como primer etapa del sistema de extracción de embocadura, previo a la etapa de clasificación.
- Salir del *bag of frames*, para utilizar la redundancia temporal y la información a priori del problema.

## Anexo: Representación gráfica de la extracción de embocadura

A continuación se muestra gráficamente la predicción del clasificador *K-Nearest Neighbors* con las características *Mel-Frequency Cepstral Coefficients* con ventana de análisis de  $23ms$  y 20 coeficientes computados con 40 bandas Mel. Para la comparación se puede ver también el *ground truth* representado en la mitad izquierda de los rectángulos.

La representación se genera asignando a los canales *RGB* de una imagen a color, las probabilidades correspondientes de cada clase a la salida del clasificador. Asignando la probabilidad de *Blow Hole Covert* al canal rojo, *Breathy Embouchure* al verde y el restante *Normal Embouchure* al canal azul se logra el resultado de la Figura 9.

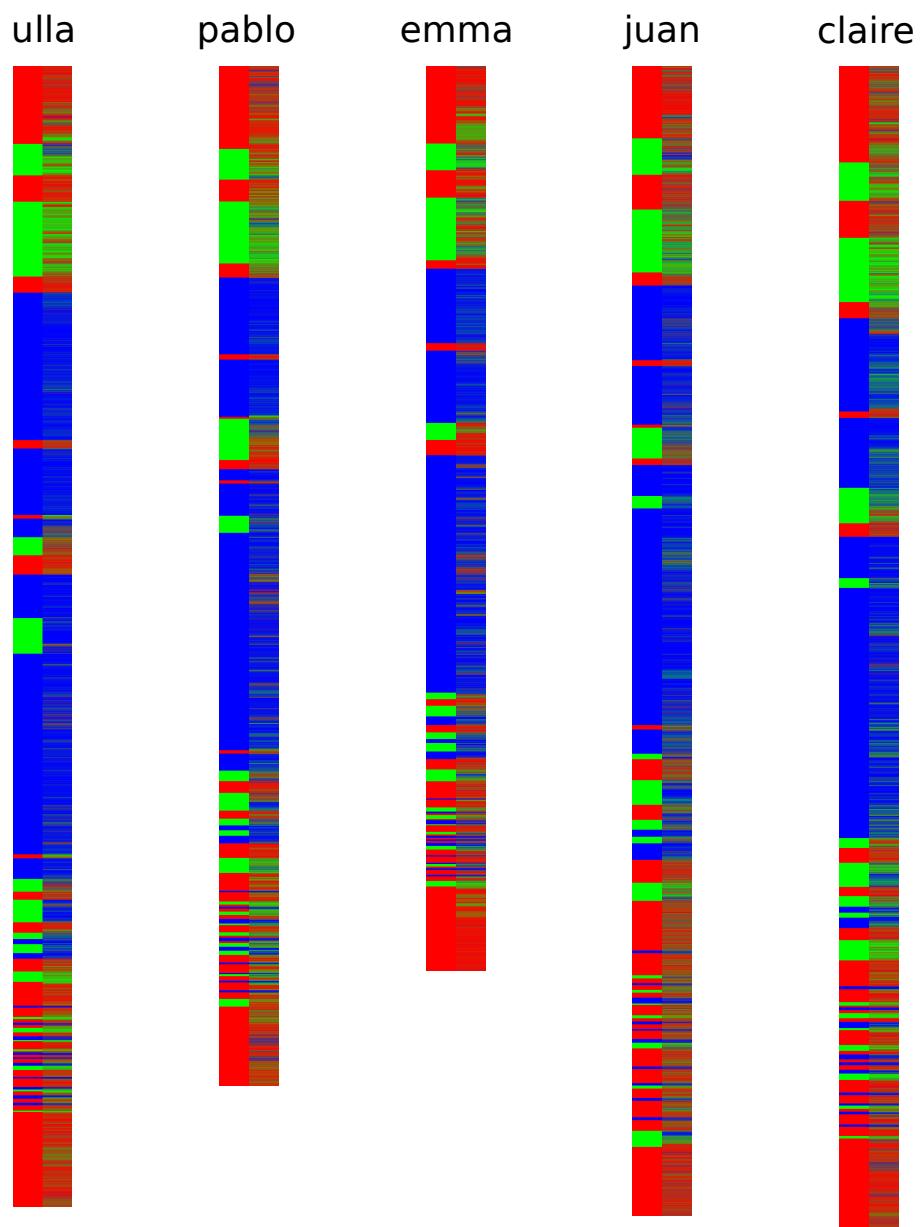


Figura 9: Representación gráfica de la predicción Vs. el *ground truth*.

## Referencias

- Candelaria, L., Costa-Gomi, E., and Hughes, P. (2005). Argentine music for flute with the employment of extended techniques: an analysis of selected works by eduardo bertola and marcelo toledo.
- Cannam, C., Landone, C., and Sandler, M. (2010). Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1467–1468. ACM.
- Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366.
- De Cheveigné, A. and Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930.
- Dick, R. (1975). *The other flute: a performance manual of contemporary techniques*. Oxford University Press.
- Jiang, D.-N., Lu, L., Zhang, H.-J., Tao, J.-H., and Cai, L.-H. (2002). Music type classification by spectral contrast feature. In *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, volume 1, pages 113–116. IEEE.
- Klapuri, A. and Davy, M. (2007). *Signal processing methods for music transcription*. Springer Science & Business Media.
- Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580.
- McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.
- Piston, W. (1955). *Orchestration*. Norton.
- Quatieri, T. (2002). *Discrete-time Speech Signal Processing: Principles and Practice*. Prentice-Hall signal processing series. Prentice Hall PTR.
- Rabiner, L. and Schafer, R. (1978). *Digital Processing of Speech Signals*. Prentice-Hall signal processing series. Prentice-Hall.
- Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190.