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

10 de diciembre de 2016

1. Introducción

1.1. La Flauta Traversa

Técnicas extendidas y técnicas tradicionales papurri

1.2. Embocadura

El término embocadura refiere al aparato de producción de la exitación de la columna de aire, en conjunto a la técnica de soplido (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 puntulaes, 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 intéprete. 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.

1.3. 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 insturmento 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 el ángulo de la flauta (ver 1.2), 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 períodicos 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.

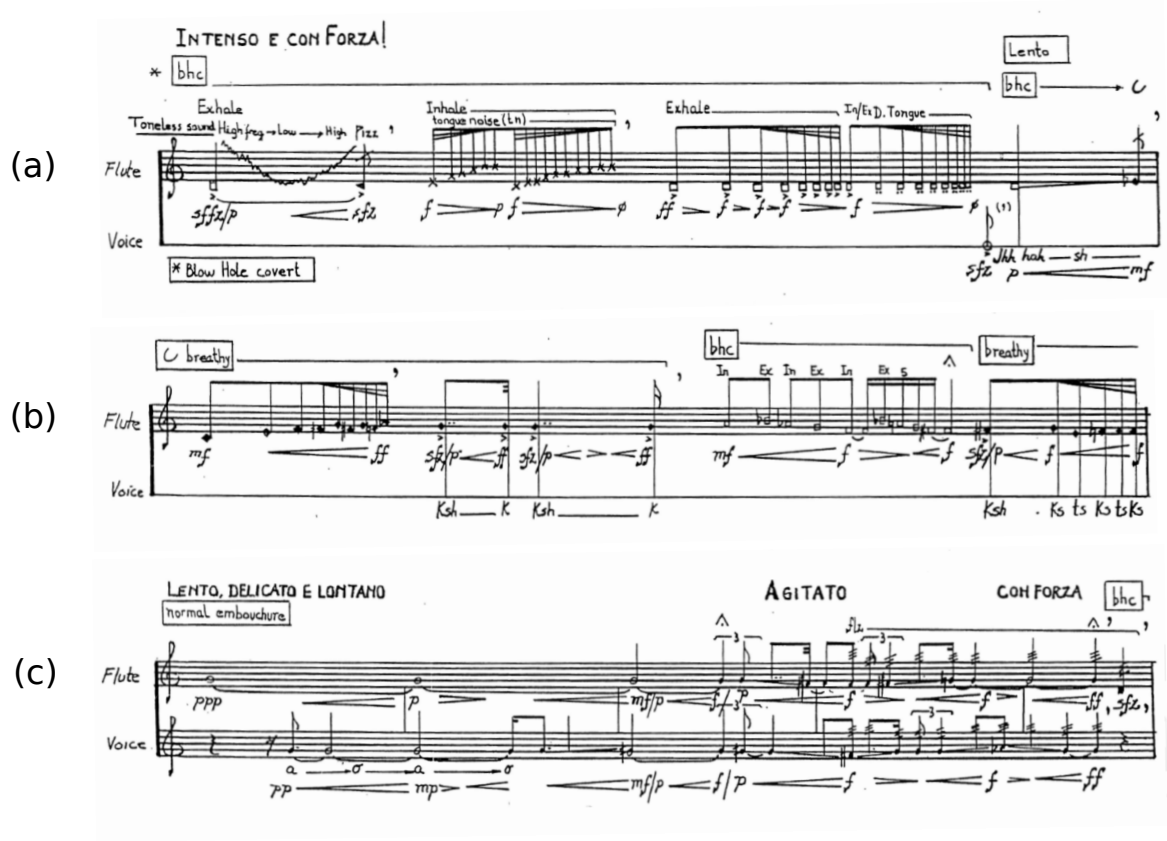

Figura 1: Notación de las embocaduras se observa en la parte superior de los sistemas. (a) *Blow Hole Covert*. (b) *Breathy Embouchure*. (c) *Normal Embouchure*. Fragmentos extraídos de la partitura de Aliento/Arrugas.

2. Definición del Problema

Teniendo en cuenta que la embocadura es un elemento determinante del material sonoro ejecutado perceptible de forma auditiva (Sección 1.2), 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 demoninación 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).

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).

MOSTRAR LAS PROPORCIONES DE LAS CLASES! NO SEAS MALO!

3. Análisis físico de la producción de sonido en la flauta

Primer experimento: con Voicing y ZCR (Rabiner and Schafer, 1978, Capítulo 4) El compositor Color Material sonoro Embocadura

4. Experimentos y Resultados

Se evalúa el desempeño de características de diversa naturaleza en la extracción del tipo de embocadura. 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.

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.

4.1. Extración de características

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

4.1.1. Características Clásicas

4.1.2. Mel-Frequency Cepstral Coefficients (MFCC) (Davis and Mermelstein, 1980)

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. 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 decorrleación que presentan los datos en el dominio de las *quefrencys*, 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 *quefrencys*, luego de la Transformada Coseno. Determina la cantidad de coeficientes, o en otras palabaras la dimensión del espacio de caracterísitcas.

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 y espectral. 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.

4.1.3. Linear Prediction Coefficients (LPC)

4.1.4. Spectral Contrast (Jiang et al., 2002)

4.2. Resultados

PRIMERO MOSTRAR COMPARACIÓN ENTRE FEATURES!!!! Y ME QUEDO CON MFCC PORQUE ES UN FIERRO PRESENTA DOS PROBLEMÁTICAS: NINGÚN FEATURE LOGRA SEPARAR BREATHY VS. BHC: se prueba solo con esas dos clases y mfcc20 (por simplicidad para el computo y clasificador) -LA CLASE TONAL QUE SERÍA LA MÁS SEPARABLE PRESENTA CONFUSIÓN QUE PODRÍA EVITARSE CON EL VOICING (?) : MFCC es un fierro pero apriori no tiene porque utilizar la información de periodicidad propiamente dicha. Parámetro relevante en la separación de estos tres tipos de embocadura. Se agrega voicing tipo yin De Cheveigné and Kawahara (2002)

4.3. Experimentos 2: Refinamiento MFCC

Compromiso resolución en frecuencia resolución temporal
Se identifica nuevamente el problema en la separación
Comparar confusión entre 2x2 y 3x3 en la separación de BHC y BREATHY
Discusión de agregar voicing, ver como mejora la matriz de confusión

4.3.1.

5. Trabajo a futuro

- Segmentación de audio, umbralizando los silencios y las respiraciones
- Salir del bag of frames, para utilizar la redundancia temporal
- Dejar Planteado el sistema completo! diagrama de bloques!

Referencias

Candelaria, L., Costa-Giomi, 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.

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.