Universidad de la República

Facultad de Ingeniería

Procesamiento Digital de Señales de Audio

Proyecto Final | Dueto de Voces Automático

Grupo 5:
Federico Bello
Juan Pablo Carballal
Juan Pedro Maestrone

28 de septiembre de 2025

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Índice

Introducción

La música es una expresión artística única que puede ser interpretada de múltiples formas por diferentes artistas. Cada versión de una canción tiene su propia identidad, matices y características especiales que la distinguen de otras interpretaciones.

Este proyecto explora las posibilidades tecnológicas de combinar dos versiones diferentes de una misma canción, utilizando técnicas de procesamiento de señales de audio.

Objetivo

En el proyecto se trabajó con dos versiones de una misma canción, interpretadas por dos artistas distintos. El objetivo principal consiste en manipular ambas interpretaciones de manera automática para obtener un dueto de voces. Esto es, integrar las versiones de manera que parezcan estar interpretando en una misma pieza musical en conjunto.

Parte 1 - Estimar la afinación de cada una de las canciones

En esta parte se calculó la afinación de los audios mediante la función "estimate_tuning" de la biblioteca librosa. La misma calcula la diferencia de afinación del audio respecto a un tono de referencia estándar de 440 Hz.

Los resultados obtenidos fueron:

- Desviación de la afinación de "The Police": 0.09 semitonos.
- Desviación de la afinación de Emily: 0.31 semitonos.

Parte 2 - Plantear dos representaciones que capturen las características de la melodía o altura

Los chromas son representaciones espectrales de señales de audio que capturan la intensidad de las 12 clases de tonos (notas musicales) dentro de una octava, independientemente de la altura tonal. Esta representación se basa en el hecho de que en la música, las notas que pertenecen a la misma clase de tono suelen desempeñar un papel similar desde un punto de vista armónico y melódico.

En este documento se comparan tres implementaciones distintas de cromagramas que ofrece la biblioteca librosa:

- `chroma_stft`: En esta implementación, el cromagrama se calcula utilizando la Transformada Rápida de Fourier (STFT) de la señal.

  Tiene las ventajas de ser rápido y eficiente de calcular, además de capturar adecuadamente las características armónicas en señales limpias y sin ruido. Sin embargo, presenta desventajas como su poca robustez frente a señales con altos niveles de ruido o diferencias tímbricas significativas, y puede generar resultados inconsistentes cuando las señales tienen características espectrales muy diferentes.

- `chroma_cqt`: Esta implementación utiliza la Transformada de Fourier Constante-Q (CQT) para calcular el cromagrama.

  Tiene las ventajas de ofrecer una mejor resolución en las frecuencias bajas y ser más preciso en la representación de estructuras armónicas que `chroma_stft`. Sin embargo, presenta desventajas como ser menos robusto frente a diferencias tímbricas pronunciadas y requerir más procesamiento en comparación con `chroma_stft`.

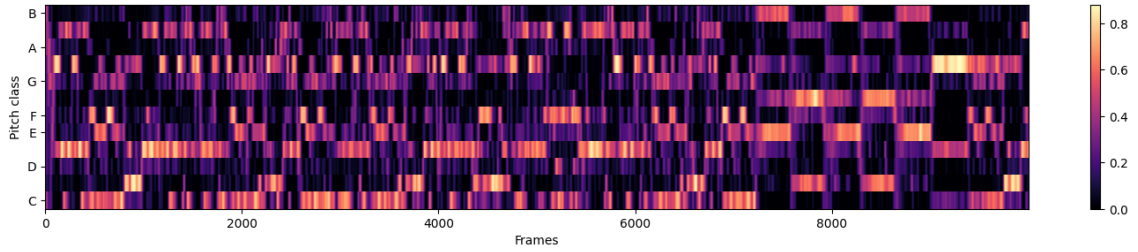
- `chroma_cens`: Esta implementación utiliza una representación cromática avanzada basada en cuantización de energía, normalización y suavizado.

  Tiene las ventajas de ser altamente robusto frente a variaciones tímbricas y dinámicas, y proporcionar una representación estable, incluso cuando las señales presentan diferencias significativas en su composición espectral. Sin embargo, presenta la desventaja de que puede perder información dinámica debido al proceso de suavizado.

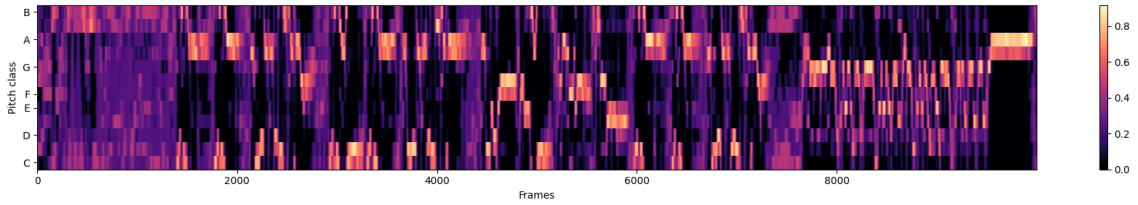
En este caso, las propiedades de `chroma_cens` lo convierten en la mejor opción para alinear las dos versiones, ya que puede capturar las similitudes armónicas independientemente de las diferencias tímbricas y dinámicas. El resultado de utilizar este chroma en ambas versiones de la canción se observa en la Figura 1, mientras que el de aplicar el chroma STFT en la Figura 2.

Previo al cálculo de los chromas, es fundamental asegurarse de que ambas versiones de la canción estén afinadas en el mismo tono. Esto implica corregir cualquier desviación de frecuencia con respecto a la afinación estándar (A440). Para este propósito, se utilizó la función `estimate_tuning`, que permitió estimar la desviación tonal de cada versión. Esta información se incorporó mediante el parámetro `tuning` al calcular los cromagramas, garantizando que ambas versiones estuvieran afinadas al mismo tono y haciendo posible una comparación precisa.

Adicionalmente, se consideraron las diferencias tonales entre las versiones. En particular, la versión de *Emily Linge* está afinada dos tonos por debajo de la versión original. Para compensar esta diferencia y alinear las tonalidades, se desplazaron los valores del cromagrama de esta versión hacia abajo en dos tonos utilizando un desplazamiento circular con la función `np.roll`. Este ajuste final permitió comparar directamente las características armónicas y melódicas de ambas versiones.

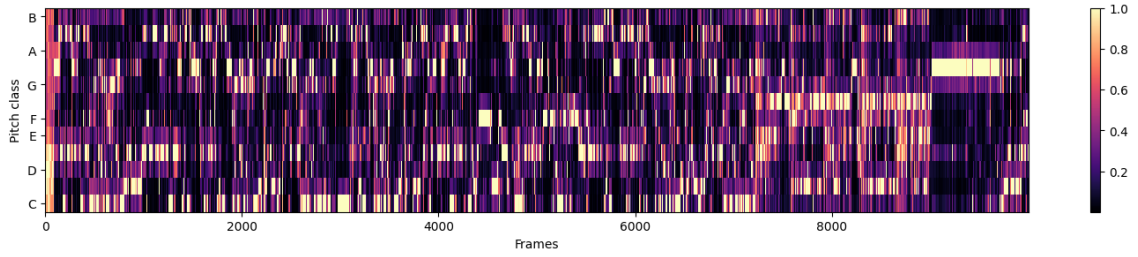


(a) Versión de *The Police*.

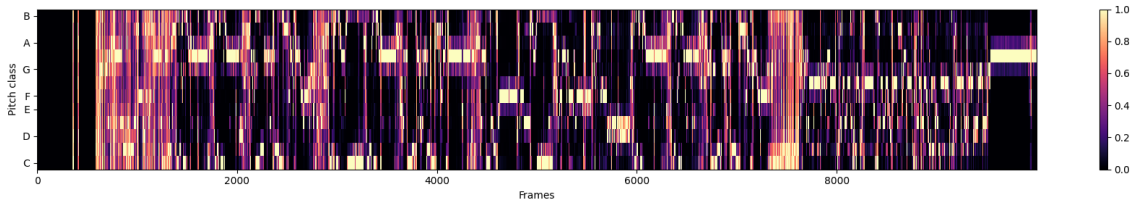


(b) Versión de *Emily Linge*.

Figura 1: Chromas CENS para las distintas versiones.



(a) Versión de *The Police*.



(b) Versión de *Emily Linge*.

Figura 2: Chromas STFT para las distintas versiones.

Parte 3 - Correspondencia temporal

Dynamic Time Warping

El algoritmo Dynamic Time Warping (DTW) es una técnica ampliamente utilizada para medir la similitud entre dos secuencias temporales que pueden variar en velocidad o duración. Este método es particularmente útil en problemas donde las secuencias a comparar no están perfectamente alineadas en el tiempo, como señales de audio, datos biométricos, patrones de movimiento o series temporales en general.

El DTW calcula una distancia entre dos secuencias al encontrar la alineación óptima que minimiza una métrica de disimilitud, generalmente la distancia euclidiana o coseno, entre sus elementos. A diferencia de una comparación directa (como la distancia euclidiana estándar), el DTW permite "estirar" o "comprimir" las secuencias en el eje temporal para encontrar una correspondencia que sea significativa.

Dadas dos secuencias temporales, $X = \{x_1, x_2, \ldots, x_n\}$ e $Y = \{y_1, y_2, \ldots, cy_m\}$, el DTW encuentra la alineación $W = \{(i,j)_k\}_{k=1}^{K}$, donde i y j son índices de las secuencias X y Y, respectivamente, y K es la longitud del camino de alineación. La alineación W debe cumplir las siguientes condiciones:

- **Condición de Monotonicidad:** El índice en X e Y debe ser no decreciente:

$$i_k \leq i_{k+1} \quad \text{y} \quad j_k \leq j_{k+1}.$$

- **Condición de Continuidad:** No se pueden omitir elementos en la alineación; cada punto de una secuencia debe alinearse con al menos un punto de la otra.
- **Condición de Frontera:** El alineamiento debe comenzar en $(1,1)$ y terminar en (n,m).

El objetivo del DTW es encontrar el camino W que minimice el costo total de alineación, definido como la suma de las distancias entre los puntos correspondientes:

$$DTW(X,Y) = \min_{W} \sum_{k=1}^{K} d(x_{i_k}, y_{j_k}),$$

donde $d(x_i, y_j)$ es la métrica de distancia antes mencionada.

En el caso particular del audio, el algoritmo suele ser aplicado sobre características *cromáticas* de la señal. Los *chromas*, o características cromáticas, son una representación compacta que captura la energía en cada una de las 12 clases de tonos de la escala cromática musical, independientemente de la octava. Esto permite extraer información que es directamente relevante para tareas musicales, dejando de lado detalles que no aportan al análisis como puede ser el *pitch*.

Uso de chromas en DTW

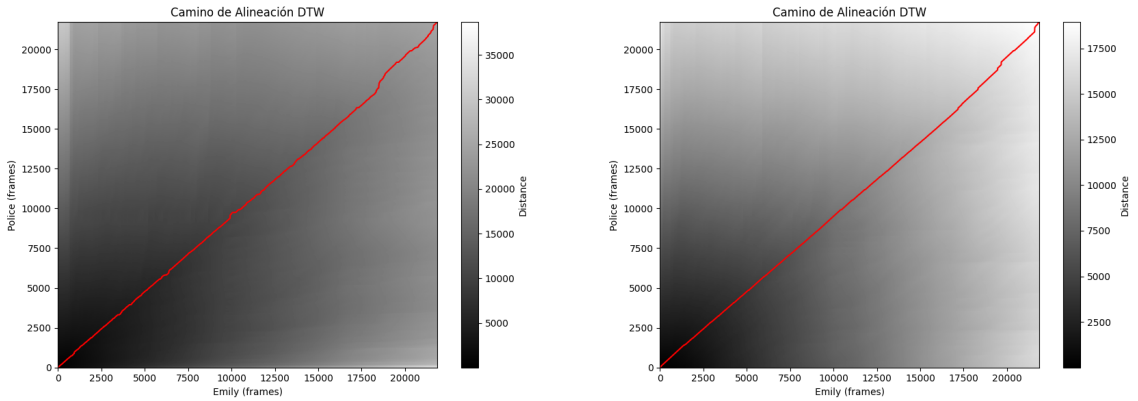
El uso de chromas en DTW es especialmente ventajoso debido a las siguientes razones:

- **Reducción dimensional**: En lugar de trabajar con la señal de audio completa o con representaciones espectrales de alta resolución, los chromas proporcionan una descripción compacta que captura la información armónica esencial. Esto facilita el cálculo y reduce la complejidad computacional del DTW.
- **Robustez a variaciones en timbre**: Al enfocarse únicamente en las propiedades armónicas, los chromas son menos sensibles a diferencias de timbre o instrumentación, lo que permite comparar secuencias musicales interpretadas en diferentes instrumentos o por diferentes intérpretes.
- **Invariancia tonal**: Como los chromas representan las notas independientemente de su altura, permiten comparar secuencias que han sido transpuestas en tono.

La aplicación del algoritmo, empleando los diversos chromas, se presenta en la Figura 3. Se observa que la alineación óptima en ambos escenarios se aproxima a la diagonal de la matriz de

costos. Esto sugiere dos aspectos: primero, en este caso específico, el algoritmo muestra poca sensibilidad respecto a la elección del chroma; segundo, las canciones están principalmente alineadas, requiriendo solo ligeros ajustes de retraso o adelanto en ciertas partes.

Este algoritmo se aplicó mediante la implementación de librosa. Esta implementación tiene el parámetro "metric", que representa la métrica de distancia mencionada en la explicación del algoritmo. Para este caso en particular, se ajustó dicho parámetro validando auditivamente los resultados de la parte 4 del proyecto, concluyendo que con el parámetro "metric = euclidean", el ajuste entre los audios tiene un resultado satisfactorio.



(a) Chroma STFT.

(b) Chroma CENS.

Figura 3: Caminos de alineación DTW para distintos chromas con distancia euclidiana.

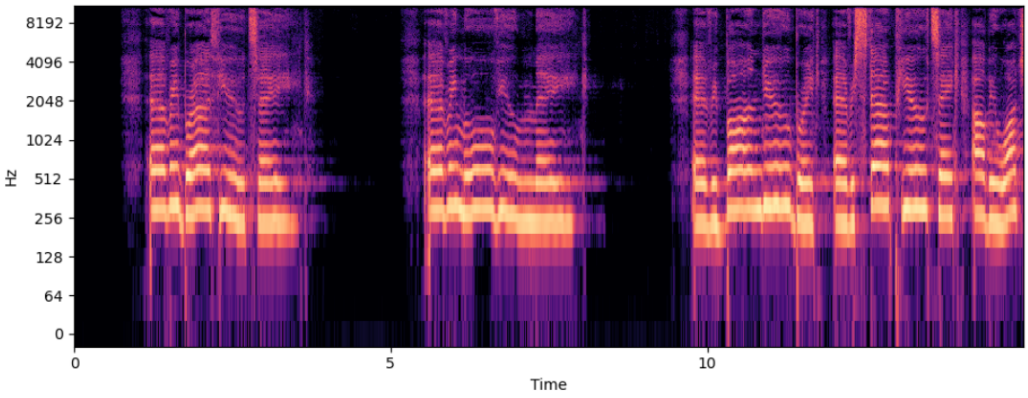
Finalmente, al combinar las canciones y escucharlas simultáneamente, el resultado auditivo, aunque presenta un carácter algo "'robótico" y con transiciones ligeramente abruptas, muestra claramente cómo las voces logran çantar al unísono". Esto evidencia que el proceso de alineación fue exitoso, cumpliendo su objetivo principal de sincronizar las señales de manera coherente en el tiempo. Este resultado destaca la eficacia del DTW para ajustar secuencias con desfasajes significativos sin sacrificar la consistencia global.

Parte 4 - Alineación Temporal

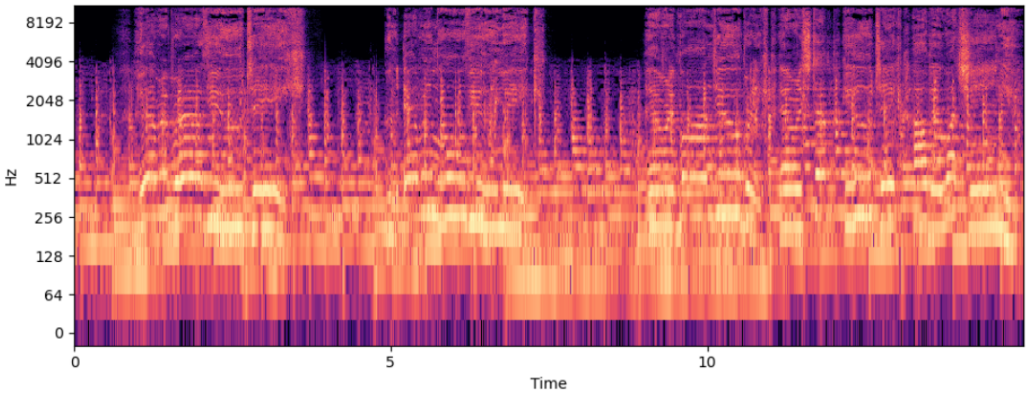
En esta parte del proyecto, se decidió utilizar el phase vocoder como herramienta principal para realizar la alineación temporal entre las canciones. Esta técnica permite modificar de manera precisa la duración de un audio sin alterar su altura tonal, lo cual resulta ideal para ajustar las diferencias temporales entre dos señales mientras se preservan sus características espectrales. La metodología se complementó con los resultados obtenidos del DTW en las partes anteriores para determinar la correspondencia óptima entre las ventanas de ambas canciones.

Siguiendo el enfoque de implementación del phase vocoder visto en el curso, se calcularon los frames de la señal de Emily de manera individual. Utilizando los resultados obtenidos a través del algoritmo DTW, se determinaron los índices de los frames correspondientes de *The Police* para cada frame de Emily. Finalmente, empleando la técnica de *overlap-add*, se reconstruyó la señal alineada de Emily, combinando los frames ajustados en el dominio del tiempo. Este proceso permitió lograr una transición suave entre los frames y preservar la continuidad de la señal final.

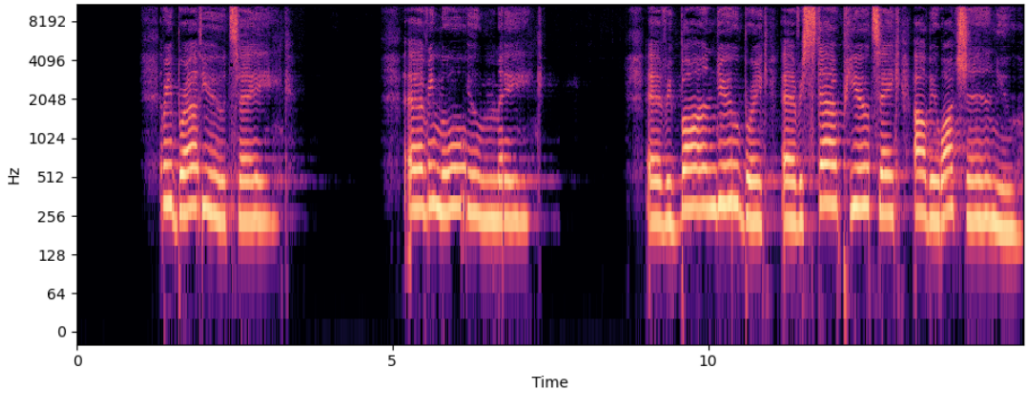
En la Figura 4, se muestran los espectrogramas de Emily sin alineación y alineados, ubicados arriba y abajo del espectrograma de "The Police", respectivamente, entre los segundos 15 y 30. Se observa que la versión alineada guarda una mayor correspondencia con la de "The Police".



(a) Espectrograma de Emily sin alinear.



(b) Espectrograma de “The Police”.

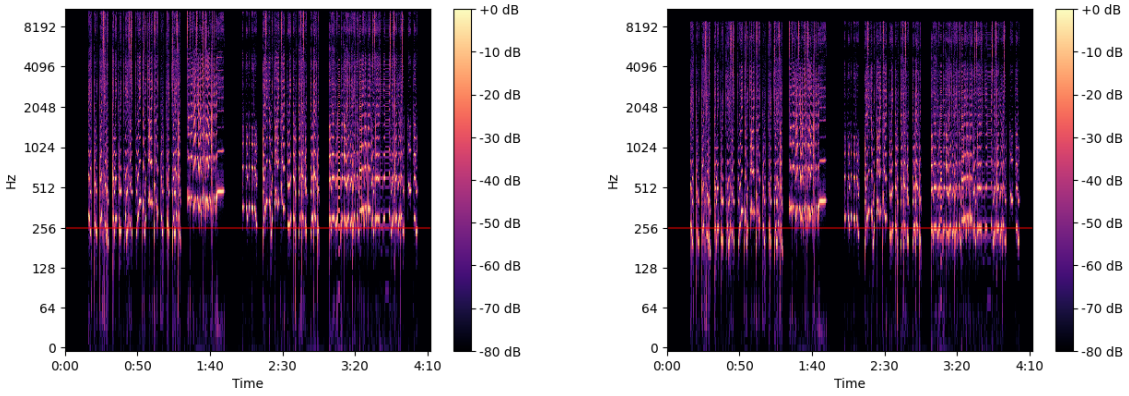


(c) Espectrograma de Emily alineado.

Figura 4: Espectrogramas de los audios desde el segundo 15 al segundo 30.

Parte 5 - Ajuste de Altura

Para ajustar la altura, se comenzó estimando el *pitch* promedio de cada señal de audio. Posteriormente, se empleó la función *pitch_ shift* de *librosa* para realizar el ajuste. Los resultados de este proceso se ilustran en la Figura 5. La Figura 5b muestra el *pitch* reducido en 2.89 semitonos en comparación con la Figura 5a.



(a) Audio original.

(b) Audio con altura ajustada.

Figura 5: Espectrograma del audio de Emily Linge con y sin ajuste de altura (notar la diferencia de altura comparando con la recta a 256 Hz).

Parte 6 - Envolvente espectral

La envolvente espectral describe cómo se distribuye la energía de una señal de audio a través de diversas frecuencias, reflejando las cualidades tímbricas del sonido. Al modificar la altura de una señal de audio, no solo se altera su frecuencia fundamental, sino que también se modifican sus cualidades espectrales, lo que puede provocar distorsiones perceptibles en el timbre original.

Para conservar la envolvente inicial, se realiza el siguiente procedimiento:

- Se calculan las STFT de la señal original y de la señal modificada.
- Se estiman las envolventes espectrales de ambas STFT mediante la función librosa.lpc.
- Se ajusta la envolvente de la STFT del audio modificado multiplicando por un factor de la siguiente forma: $\text{factor} = \dfrac{\text{envolvente\_original}}{\text{envolvente\_modificada} + \varepsilon}$
- Finalmente, el proceso concluye aplicando la ISTFT.

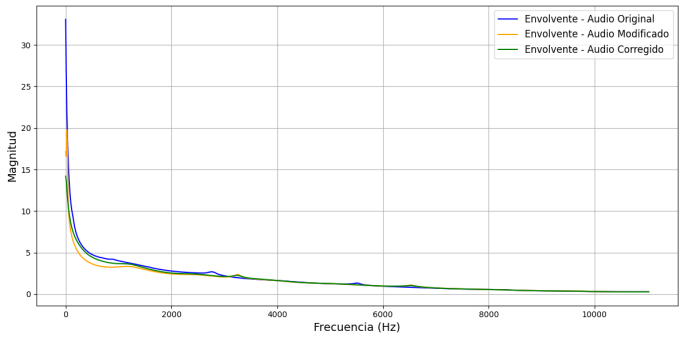


Figura 6: Comparación de las envolventes espectrales para el audio de Emily Linge: el audio original alineado, el audio modificado con cambio de altura y el audio corregido, donde se aplicó un ajuste para preservar la envolvente espectral utilizando LPC.

En la Figura 6, se observa que la envolvente del audio corregido se aproxima considerablemente más a la del audio original, especialmente en las frecuencias bajas y medias. Esto indica que el método de ajuste mediante LPC ha sido efectivo para alinear la envolvente espectral del audio modificado con la del original, manteniendo mejor las características espectrales principales. Sin embargo, en las frecuencias altas, las diferencias entre las tres envolventes son menores, lo que sugiere que el ajuste afecta principalmente las regiones de baja y media frecuencia, donde la energía es más significativa.

Parte 7 - Representación basada en características fonéticas

Para preparar los audios utilizados en esta etapa, se realizó una separación de las fuentes de señal en la versión original de *Every Breath You Take* de The Police, aislando la señal vocal de la instrumental. Esta separación se llevó a cabo utilizando la aplicación `Spleeter` de *Deezer*, disponible en GitHub, y también mediante un notebook en Google Colab, que permite ejecutarla sin necesidad de instalaciones adicionales. Este procedimiento se aplicó tanto para esta sección como para la parte 8 del documento.

Para esta parte se llevó a cabo la alineación temporal utilizando como base la **envolvente espectral** calculada a partir de ventanas de tiempo a lo largo del archivo de audio. La envolvente espectral actúa como una representación aproximada basada en fonemas, ya que captura las características acústicas generales del habla, particularmente los **formantes**, que son elementos clave para identificar los fonemas. Al derivar la envolvente espectral de cada ventana temporal, se obtiene una representación suavizada de la distribución de energía en las frecuencias, reduciendo la influencia de los detalles armónicos y centrando la descripción en la estructura espectral general, estrechamente vinculada a la articulación fonética.

Cada fonema se distingue por una configuración específica de formantes, que reflejan concentraciones de energía en rangos espectrales concretos. La envolvente espectral captura esta distribución de energía, permitiendo diferenciar entre diversos sonidos vocálicos y consonánticos.

Esta representación es especialmente útil para alinear interpretaciones vocales que, aunque puedan variar considerablemente en aspectos melódicos, conservan una estructura fonética similar. Los resultados obtenidos al aplicar DTW con esta representación muestran un desempeño comparable al logrado con características cromáticas, lo que confirma que la alineación puede realizarse de manera efectiva.

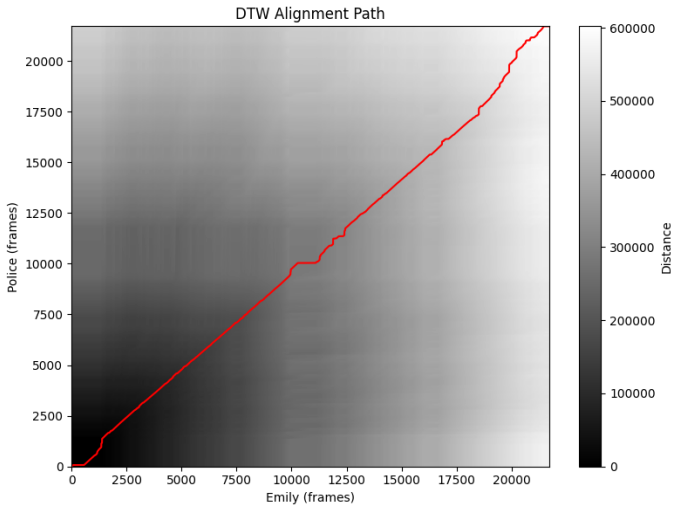


Figura 7: Alineación entre los audios utilizando la envolvente espectral.

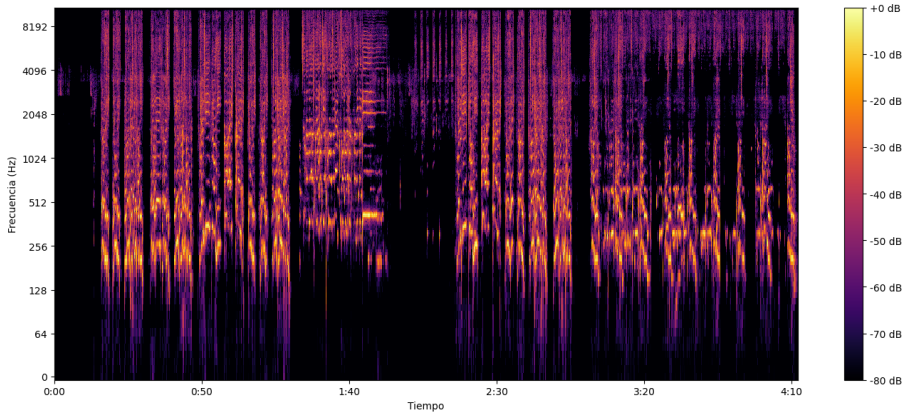
La Figura 7 muestra el camino de alineación obtenido mediante DTW con la envolvente espectral. Al compararla con las alineaciones basadas en cromagramas (Figura 3), se aprecia que ambas métricas generan caminos similares en la matriz de costos.

La Figura 8 ilustra los espectrogramas del audio procesado, mostrando, por un lado, la señal de "The Police" sin el instrumental de fondo y, por otro, la versión de Emily Linge alineada. Se destaca la alineación a nivel temporal entre ambas señales.

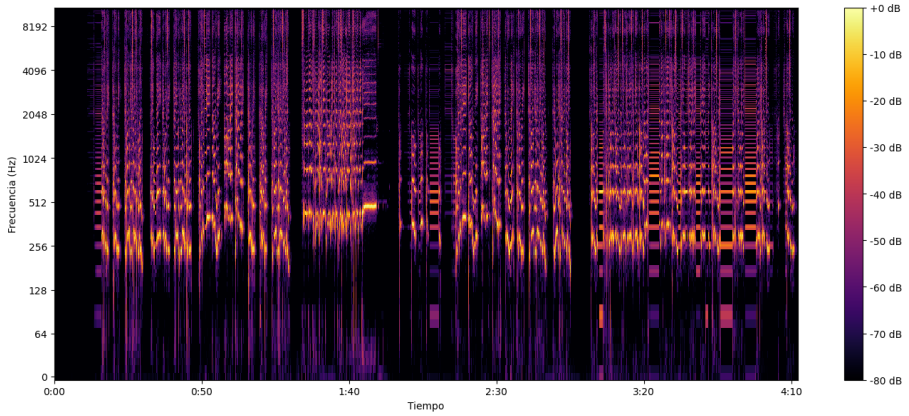
Parte 8 - Alternado de las voces

El objetivo de este proceso es crear un efecto de dueto alternando las voces de dos artistas en los versos de una canción, de manera que parezca que ambos están interpretando la pieza de forma conjunta y coordinada.

El procedimiento llevado a cabo se puede describir de la siguiente manera:



(a) Espectrograma del audio de The Police sin instrumental de fondo.



(b) Espectrograma del audio de Emily Linge alineado.

Figura 8: Comparación de espectrogramas entre el audio de The Police sin instrumental y el audio de Emily Linge luego de alinearlo con la envolvente espectral.

1. **Detección de versos:** Se identifican los segmentos de la señal de audio donde hay actividad vocal. Para esto, la señal se divide en ventanas temporales, calculando la energía en cada una. Si la energía de una ventana supera un umbral predeterminado, se considera que hay actividad vocal. Las ventanas consecutivas con actividad se agrupan en intervalos de tiempo, descartando aquellos intervalos cuya duración sea demasiado corta para evitar incluir ruido o silencios accidentales.

2. **Silenciamiento alternado:** Una vez detectados los versos, se procesan las dos versiones de la canción de forma independiente:

   - En la primera versión, se silencian los versos impares.
   - En la segunda versión, se silencian los versos pares.

3. **Combinación de señales:** Finalmente, se combinan las señales procesadas de ambas versiones sumándolas. Esto genera una señal resultante en la que las voces de los dos artistas alternan durante los versos, logrando el efecto de dueto buscado.

El resultado de este proceso es una interpretación alternada en la que las voces de los artistas parecen complementarse, creando una experiencia auditiva similar a la de un dueto. La Figura 9 ilustra este proceso, mostrando ambas señales superpuestas con segmentos silenciados en colores distintos.

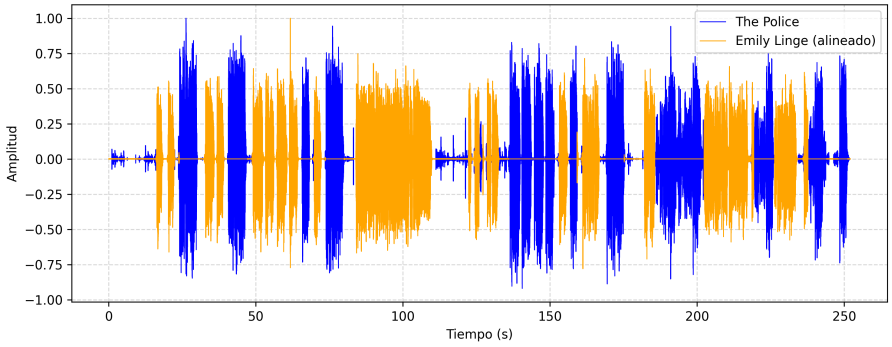


Figura 9: Audio resultante de superponer las dos señales de audio, silenciando "The Police" en los versos impares, y "Emily Linge" en los pares.

Parte 9 - Armonizado en el estribillo

El procedimiento realizado para generar una armonización con desplazamientos tonales se describe a continuación:

1. **Cambio de tonalidad:** En una primera instancia, utilizando el efecto de transposición en frecuencia implementado en el curso, se desplazó la altura tonal del audio de *Emily* y se sumó a la canción de *The Police*, aplicándolo exclusivamente durante los estribillos de la canción.

2. **Conservación de la duración:** Para modificar la tonalidad del audio sin alterar su duración:

   - Se aplicó un estiramiento temporal mediante la implementación de *phase vocoder* de la biblioteca `librosa`. Este estiramiento utiliza una razón proporcional a $2^{\frac{\#\text{semitonos}}{12}}$.
   - Posteriormente, la señal estirada fue remuestreada utilizando la función `resample_poly` de `scipy`, ajustando la duración para que coincida con la original.

   Para optimizar el rendimiento del remuestreo, la razón $2^{\frac{\#\text{semitonos}}{12}}$ se aproximó a fracciones de números enteros pequeños, como se detalla en la Tabla 1.

3. **Armonización:** Se realizó la armonización para desplazamientos de 1, 3, 5 y 7 semitonos, empleando las aproximaciones indicadas en la tabla mencionada. Estas señales desplazadas y ajustadas se normalizaron y sumaron a la pista base únicamente durante los intervalos correspondientes a los estribillos, manteniendo la coherencia tonal y temporal.

Semitonos	**Razón Exacta**	**Aproximación**
7	$2^{\frac{7}{12}} \approx 1{,}498$	$\frac{3}{2} = 1{,}500$
5	$2^{\frac{5}{12}} \approx 1{,}335$	$\frac{4}{3} \approx 1{,}333$
3	$2^{\frac{3}{12}} \approx 1{,}189$	$\frac{19}{16} \approx 1{,}187$
1	$2^{\frac{1}{12}} \approx 1{,}059$	$\frac{18}{17} \approx 1{,}059$

Cuadro 1: Razones exactas y aproximaciones para diferentes desplazamientos en semitonos.

Se analizaron los resultados mediante espectrogramas para evaluar las diferencias entre la señal alineada y procesada de *Emily Linge* y la señal armonizada con un desplazamiento tonal de 5 semitonos en el estribillo. Dichos espectrogramas están visibles en las Figuras 10 y 11.

Al comparar las Figuras 10 y 11, se observa que el espectrograma de la señal sin armonizar presenta componentes de frecuencia que superan los 10 kHz, mientras que en la señal armonizada estas se limitan alrededor de los 8 kHz. Este comportamiento puede atribuirse al proceso de transposición tonal y remuestreo, que afecta las frecuencias más altas, reduciendo la riqueza armónica en esa región del espectro.

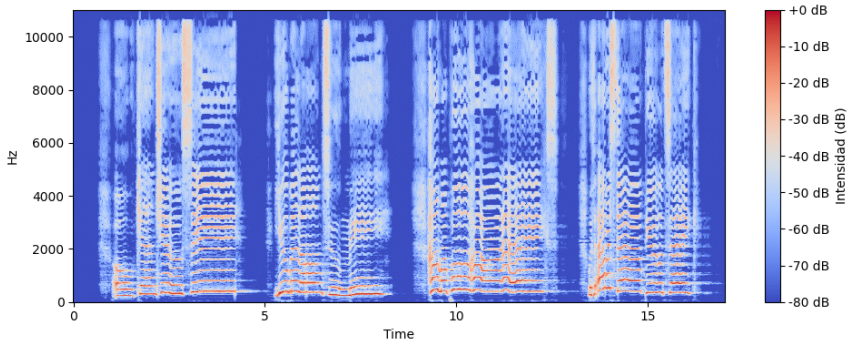


Figura 10: Espectrograma del audio de *Emily Linge* alineado y procesado durante el estribillo.

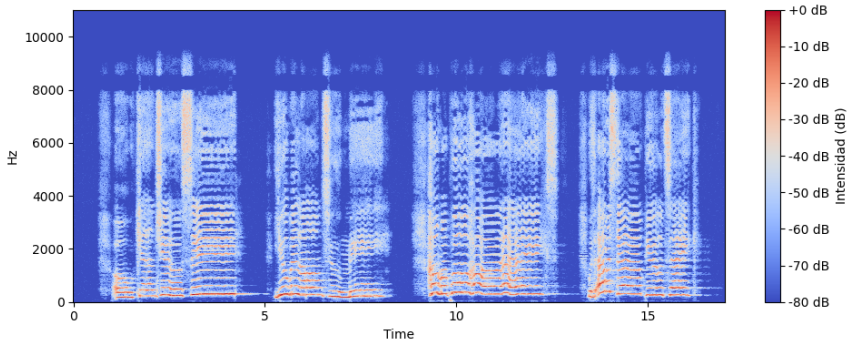


Figura 11: Espectrograma del audio de *Emily Linge* alineado, procesado y armonizado con un desplazamiento de 5 semitonos hacia abajo durante el estribillo.

Conclusiones

En este proyecto se consiguió alcanzar todos los objetivos principales, demostrando en el proceso cómo diversas técnicas de procesamiento digital de señales pueden combinarse eficazmente para integrar dos interpretaciones independientes de una misma canción y generar un dueto automático. Los resultados validan el uso de métodos como la estimación y ajuste de afinación, junto con la alineación temporal mediante *Dynamic Time Warping* (DTW), para sincronizar diferencias en tono y tempo entre las versiones. Representaciones cromáticas y fonéticas se utilizaron para capturar similitudes armónicas y estructurales, proporcionando herramientas complementarias para tareas de alineación y análisis musical.