UNTREF UNIVERSIDAD NACIONAL DE TRES DE FEBRERO

INGENIERÍA DE SONIDO

Estudio sobre cancelaciones en la mezcla de sonido directo para producciones audiovisuales.†

Autor: Francisco Messina

Tutor/es: Ignacio Mieza

Diego Martínez

(†) Tesis para optar al título de Ingeniero/a de Sonido

Diciembre 2020

AGRADECIMIENTOS

Muchas personas e instituciones fueron partícipes de forma directa o indirecta de mi formación académica y personal y de la finalización de la presente investigación.

En primer lugar quisiera agradecer a mis padres por su apoyo incondicional. Su amor, paciencia y compañía fueron imprescindibles en cada etapa de la carrera. Gracias por la confianza y el aliento. Esta tesis está dedicada en su totalidad a ellos.

A mi tutor de tesis Ing. Ignacio Mieza por la paciencia y la sabiduría. Siempre pude contar con una buena predisposición y excelente trato planificando la investigación y sorteando las diferentes dificultades que se presentaron a lo largo de ella. Tanto es así que, incluso en la mitad de un evento mundial tan inesperado como una pandemia, pude contar con Ignacio para seguir avanzando con mi investigación.

A mi Co-tutor Diego Martinez por brindar el tema de la tesis y poner a disposición su conocimiento y recursos en cada situación donde fue requerido.

A la Universidad Nacional de Tres de Febrero (UNTREF), a su Rector Lic. Anibal Jozami, a todo su personal docente y no docente. Por promover un espacio ideal para el desarrollo de ideas y nuevos pensamientos, y brindar a todos y cada uno de los alumnos de esta casa de altos estudios, todos los recursos que esta institución dispone.

Esta investigación no hubiera sido posible sin una formación académica acorde, por este motivo debo extender mi agradecimiento a los docentes de la carrera de Ingeniería de Sonido de la UNTREF, a su coordinador Ing. Alejandro Bidondo.

A mi novia Laura. Su amor me dio la fuerza y el impulso que necesité para terminar este proyecto. Esta tesis también es de ella.

A mi hermana Elena, a mis abueles, tíes y primes.

A Gabriel “Dino” Rosanigo, mi compañero de escritura, por sus aportes fundamentales a esta investigación.

A mis amigues de siempre y a les que generé a lo largo de los años de cursada.

INDICE DE FIGURAS

INDICE DE TABLAS

RESUMEN

Muchas veces, los editores de audio deben realizar tareas repetitivas, engorrosas y muy demandantes de tiempo, como el caso de la alineación de pistas de audio para corregir corrimientos temporales y diferencias de fase. Con la evolución del campo del procesamiento digital de señales, varias de esas tareas se han automatizado mediante software específicos.

Para la alineación temporal de dos tomas del mismo evento sonoro, en particular en la mezcla de un micrófono de caña y uno de solapa, en grabaciones de diálogo para producciones audiovisuales, las soluciones automáticas son escasas y muy recientes. Debido a la naturaleza de estas técnicas microfónicas, en la suma directa de dichas señales aparecen cancelaciones de frecuencia variable con comportamiento complejo. Para evitar estas cancelaciones, es necesario corregir la alineación temporal de los fragmentos de las señales antes de realizar la suma.

En la presente tesis se investiga el problema de corrección de la alineación temporal de señales de audio, y se presentan herramientas basadas en el procesamiento digital de las señales que abordan el problema. Los algoritmos propuestos se implementaron en un software, que fue validado utilizando múltiples señales de entrada; y además se comparó con otros software comerciales dedicados al mismo fin.

Palabras clave: “Posproducción de sonido”, “Alineación temporal”, “Producción audiovisual”, “Correlación cruzada”

ABSTRACT

Many times sound editors must perform repetitive, cumbersome and time-consuming tasks. Such is the case of audio track alignment for temporal and phase shift corrections. With the evolution in digital sound processing (DSP) software, many of these tasks were replaced.

For the task of automatically aligning two tracks from two different microphones of the same acoustic event, particularly the mixing of a lav and a boom microphone in audiovisual productions the DSP solutions are few and new. Due to the nature of this technique, the direct sum of both signals generates comb filtering with variable center frequency. To avoid this problem one of the signals must be divided into small fragments and perform individual time alignments for each fragment, before performing the sum.

In the present thesis work, the behavior of said phenomena is investigated. Also, a possible solution is presented by the use of DSP. The proposed algorithms are then implemented in a software. The functionality of said software is then evaluated using multiple input signals. The results are compared with other software dedicated to the same objectives.

Keywords: “Sound post production”, “Temporal alignment”, “Audiovisual production”, “Cross correlation”

1. INTRODUCCIÓN

1.1. FORMULACIÓN DEL PROBLEMA

En la mezcla de diálogos en producciones audiovisuales, es común que el técnico de mezcla reciba más de una pista microfónica del mismo evento sonoro. Habitualmente, se trata de una pista registrada con un micrófono "*Boom*" o de caña, y otra con un micrófono de solapa (Figura 1). En esta situación existen dos caminos posibles, elegir la pista que mejor se adapte al proyecto y trabajar exclusivamente con esa, o realizar una mezcla de ambas pistas intentando obtener lo mejor de cada una [1].

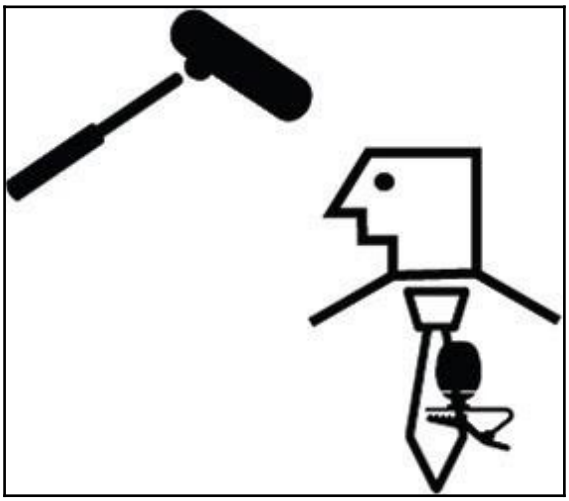

Figura 1. Técnica microfónica clásica para grabación de voz hablada en producciones audiovisuales

La primera opción suele implicar un compromiso entre calidad sonora y inteligibilidad de la voz. Los micrófonos de caña suelen captar un sonido más "natural" de la voz humana integrada en el ambiente donde se está grabando, pero son más sensibles a ruidos externos y dependen mucho de la destreza del operador y de las condiciones de toma, ya que estos micrófonos por convención se ocultan fuera del cuadro visual. Por otro lado, los micrófonos de solapa son buenos para captar con claridad los diálogos, pero entregan una grabación de voz más "artificial" [2].

En la segunda opción, si se pretende mezclar ambas señales sin una previa alineación, aparecen cancelaciones de frecuencias dentro del espectro audible. Estas cancelaciones se deben a la diferencia variable en los tiempos de arribo del sonido a cada micrófono, producto del movimiento impredecible de los mismos y de la fuente sonora; y

se pueden modelar, entre otras formas, como un filtro peine de frecuencia central variable.

Esto genera que la alineación temporal entre pistas resulte de suma complejidad, por más que inicialmente se encuentren alineadas. Históricamente, este problema se ha solucionado cortando y alineando temporalmente pequeños fragmentos de audio (a veces palabra por palabra) [3], lo que implica una tarea extremadamente engorrosa y demandante de tiempo. Por esta razón, muchos técnicos de mezcla optan por usar una única pista de audio, sacrificando calidad a cambio de reducir sustancialmente el tiempo y el esfuerzo [2].

En la actualidad, existen diversos programas comerciales dedicados a alinear pistas de audio. Entre ellos se destaca el plug-in para *Pro tools* de *Sound Radix* llamado *Auto-Align Post* [4]. Este software, cuya versión 1.0 se lanzó en Agosto del 2018, está diseñado para corregir automáticamente los filtros peine, generados por corrimientos variables en la relación de distancia entre la fuente y los micrófonos, particularmente en la situación de grabación de diálogos en vivo con más de un micrófono. Cabe destacar que, previo al lanzamiento de dicho software, no existía en el mercado, según nuestro conocimiento, ningún otro dedicado a esta tarea específica.

En la presente tesis se estudian los efectos de la suma de dos tomas microfónicas del mismo evento sonoro, en particular la grabación de diálogos. Además, se plantean las bases de un software que permite agilizar dicha tarea. El algoritmo planteado se evalúa utilizando diferentes señales de entrada, y se comparan los resultados obtenidos utilizando el *Auto-Align Post*.

1.2. OBJETIVOS

1.2.1. Objetivo general

El objetivo general de la investigación es la realización de un estudio sobre las cancelaciones producidas en la mezcla de sonido directo en producciones audiovisuales, entender sus causas y presentar soluciones automáticas basadas en algoritmos clásicos de procesamiento digital de las señales.

1.2.2. Objetivos específicos

Entre los objetivos específicos se encuentran:

- Proponer algoritmos basados en herramientas clásicas del procesamiento digital de señales que aborden el problema.
- Implementar un software básico que permita probar automáticamente los algoritmos propuestos. Para ello se utiliza el lenguaje de programación *Python*.
- Comparar el software resultante con un software comercial.
- Plantear posibles líneas de investigación a futuro.

2. ESTADO DEL ARTE

Existen numerosos procesos involucrados en la captación y edición de audio en producciones audiovisuales [5]. A continuación se detallan los principales.

2.1. GRABACIÓN

2.1.1. Micrófonos

En la grabación de diálogos para producciones audiovisuales, es común utilizar en simultáneo dos tipos de micrófono diferentes: los micrófonos de caña (*boom microphone*) y los micrófonos de solapa (*lapel microphone*). Ambos cumplen el propósito de captar el sonido del diálogo que se pretende registrar, permaneciendo invisibles ante las cámaras de video [2].

La característica principal de los micrófonos de caña es que están diseñados para ser montados sobre una caña o brazo ajustable que es manipulado por un operador especializado, y de esta manera es posible posicionar el micrófono directamente sobre el hablante [6]. Estos micrófonos son típicamente transductores a condensador con alimentación tipo phantom power, diafragma liviano y sensible, patrón polar direccional (cardioide , super o hipercardioide), respuesta relativamente plana y sin limitaciones en el tamaño de cuerpo y membrana. Opcionalmente, cuentan con un tubo de interferencia que ayuda a cancelar cualquier sonido proveniente de los laterales, generalmente utilizado en grabaciones con un alto piso de ruido. Otra característica que los vuelve óptimos para este tipo de producciones, es su bajo nivel de ruido eléctrico [7]. Dentro de la amplia variedad de marcas y modelos de micrófonos de caña en el mercado, se destacan el Sennheiser Mkh60, cuya respuesta en frecuencia y patrón polar se muestra en la Figura 2, el Neumann krm 81 y la línea CMIT de Schoepps.

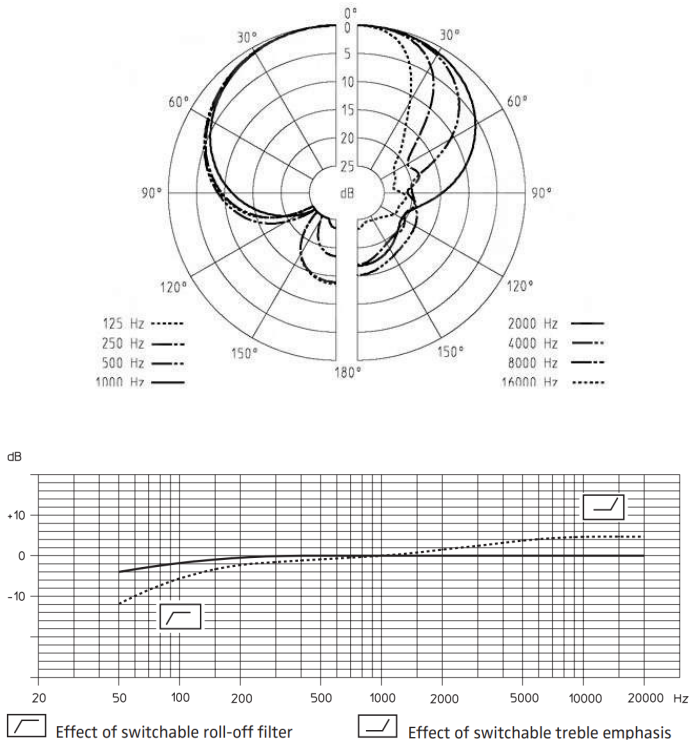

Figura 2. Patrón polar y respuesta en frecuencia del micrófono Sennheiser Mkh60 [8].

La implementación de estos micrófonos como único método de captación de diálogos puede traer problemas, principalmente debido a que su correcto uso depende, en gran medida, de la habilidad del operador y del tipo de encuadre visual. Al tratarse de micrófonos supercardioides, el mínimo corrimiento del eje de captación es capaz de modificar significativamente el timbre y la amplitud del sonido captado. También es importante destacar que los movimientos constantes propios de operar manualmente un micrófono, generan pequeñas diferencias en los tiempos de arribo del sonido.

Con el objetivo de obtener una grabación de diálogo más estable y aumentar la claridad, junto con el micrófono de caña se utiliza el micrófono de solapa. Ambos constituyen el estándar de la industria audiovisual [1]. Esto se debe a su capacidad de captar el diálogo de forma aislada y con bajo nivel de ruido. Son micrófonos electret con patrón polar omnidireccional u ocasionalmente cardioide, curva de respuesta menos lineal y a veces optimizada para la emisión lingüística, y construcción miniaturizada para facilitar posicionamiento y ocultamiento [2]. Esto genera una mayor inteligibilidad en la captación del habla. Su calidad sonora es menor que la de los micrófonos de caña, y su respuesta en bajas frecuencias depende principalmente del efecto de proximidad [9].

Dentro de las diferentes marcas y modelos de micrófonos de solapa se destaca la serie Sanken COS-11, cuya respuesta en frecuencia y patrón polar se muestran en la Figura 3. Los movimientos de la cabeza del hablante, entre otros factores, pueden variar la distancia entre la fuente y el micrófono, de la misma forma que el timbre y la amplitud del audio. Ambos micrófonos deben ubicarse, dentro de lo posible, cerca del hablante para aumentar la claridad del habla y la relación señal/ruido.

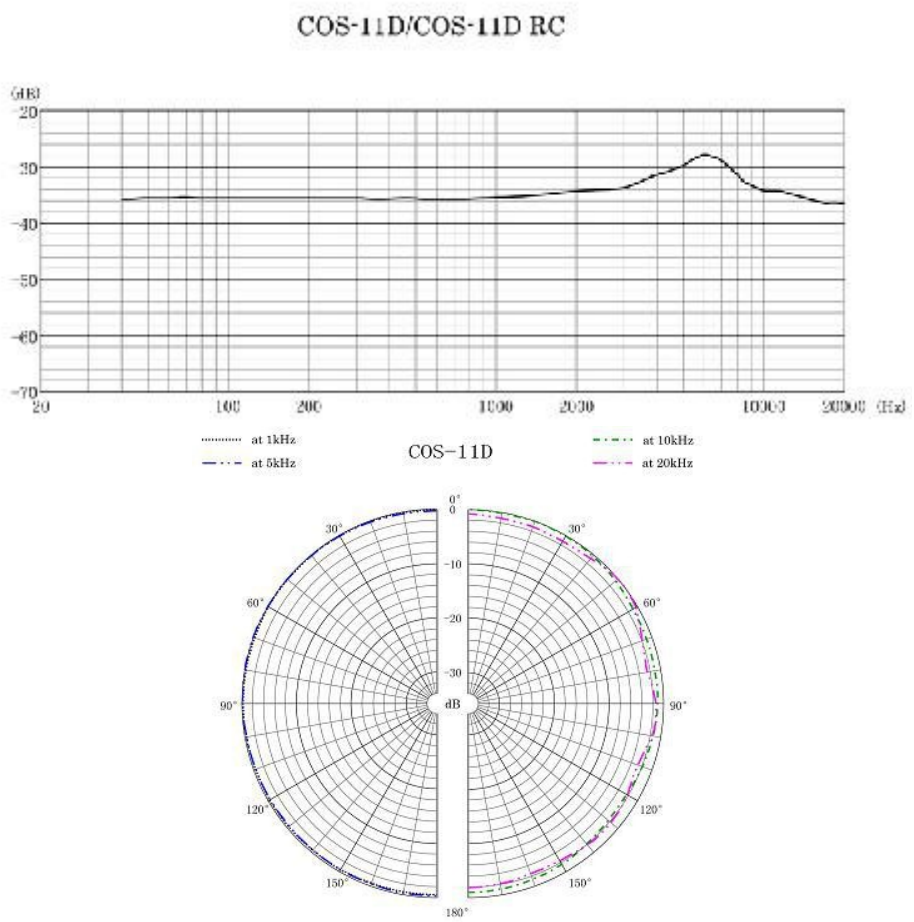

Figura 3. Respuesta en frecuencia y patrón polar del micrófono de solapa Santek COS-11D [10].

Resulta evidente que las características de ambos micrófonos son considerablemente distintas. Esto es así tanto para el espectro de las señales captadas como para su dinámica. Por un lado, los micrófonos de caña tienen un piso de ruido más alto que los micrófonos de solapa, ya que al estar a mayor distancia de la fuente captan los sonidos generados en los alrededores. Esto genera un sonido más natural de la voz humana, a la vez que pierde claridad. Por el contrario, los micrófonos de solapa tienen

una excelente relación señal a ruido lo cual genera un sonido más artificial [2]. Con respecto a la dinámica de la señal, los micrófonos de solapa son más sensibles a los movimientos de la cabeza del hablante. Esto puede generar cambios drásticos en la amplitud de la señal que no provienen de un aumento en el nivel de la fuente. Este no es el caso para los micrófonos de caña que respetan más fielmente la dinámica propia de la fuente.

2.2. MEZCLA

En la etapa de postproducción de diálogos, como primer paso se decide si se utilizará solo un micrófono (solapa o caña) o una mezcla de ambos. Esta segunda opción, como ya se dijo, permite al editor obtener las mejores características de ambos micrófonos, por un lado, la claridad que proporcionan los micrófonos de solapa y por el otro, la naturalidad y espacialidad de los micrófonos de caña [2]. Sin embargo, aparece un efecto de filtro peine generados dentro del espectro audible, debido a los corrimientos variables en los tiempos de arribo del sonido a cada micrófono [2]. Estos corrimientos variables se ven representados en la Figura 4 donde se ve una señal inicialmente en fase que pierde su alineación con el paso del tiempo.

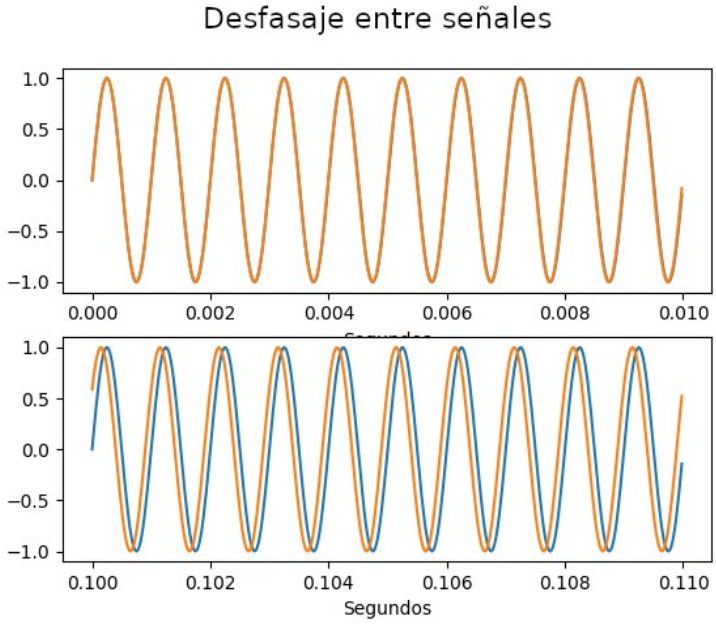

Figura 4. Desfasaje entre señales inicialmente alineadas

2.2.1. Alineación manual

Es posible solucionar este problema alineando manualmente pequeños fragmentos de ambas pistas de audio, pero esta tarea insume una enorme cantidad de tiempo para llegar a un resultado deseado. Los corrimientos entre pistas requieren en muchos casos, una alineación palabra por palabra [2]. Se recurre a la identificación en la forma de onda general de la toma, de eventos comunes, típicamente los transientes de ataque de consonantes oclusivas.

2.2.2 Alineación automática

Existen diversos software comerciales dedicados a solucionar problemas similares al planteado aquí. Dichos programas se pueden dividir en tres grupos. En el primer grupo se encuentran aquellos programas que acomodan temporalmente una pista de audio con respecto a otra de referencia, realizando una correlación global. Estos programas son muy útiles en los casos donde la relación de distancia entre ambos micrófonos y la fuente es fija. Simplemente con acomodar una de las pistas entera con respecto a la otra alcanzaría para evitar interferencias. Ejemplos de programas que cumplen con estas características son el *plug-in* de *Waves* llamado *InPhase* [11] y la herramienta *Phase* de *RX* (Izotope) [12].

En el segundo grupo se encuentran los programas que usan la deformación temporal (*Time stretch*) para acomodar una pista de audio a otra de referencia. Las pistas no necesitan ser del mismo evento sonoro, pero sí similares. El programa modifica la longitud de la pista a editar. Estos programas suelen utilizarse para simplificar el trabajo de reemplazo automático de diálogo (*ADR*), donde se debe sincronizar un video con una pista de audio grabada en postproducción. En este caso, la pista de referencia se utiliza solamente para alinear la de audio grabada en postproducción con la imagen, y no será reproducida en el producto final, por lo que no existen problemas de cancelaciones. Algunos ejemplos de estos programas son, la herramienta *Automatic Speech Alignment* del *Adobe Audition* [13] y el *PluralEyes* de *Red Giant* [14].

Ambos grupos de programas solucionan problemas similares al planteado aquí, pero resultan insuficientes por diferentes razones. Para los programas del primer grupo, es necesario que tanto los micrófonos como la fuente sean fijas. Si se introducen variaciones en la posición de alguno de los tres elementos durante la grabación, estos

programas serían incapaces de evitar la suma destructiva de señales a lo largo de toda la pista.

Por su parte, los programas del segundo grupo deforman temporalmente la señal desplazada modificando su fase, y por lo tanto resulta imposible la tarea de alinear con la pista de referencia. Otro motivo por el cual resultan insuficientes para el objetivo planteado, es su falta de precisión. Al estar orientados al *ADR*, alcanzaría con generar diferencias temporales menores a 45 ms para que el audio y el video se perciban alineados [4]. Sin embargo, en caso de tener desfasajes de 40 ms o más entre dos pistas de audio con señales similares, se percibirían cancelaciones dentro del espectro audible.

Por último, el tercer grupo de programas dedicados a solucionar problemas de alineaciones temporales entre pistas de audio está integrado únicamente por el *Auto-Align Post* de S*ound Radix*, cuya pantalla principal se muestra en la Figura 5, mencionado en la formulación del problema de la presente tesis [4]. Este programa alinea de forma automática y dinámica la fase de grabaciones de diálogo capturada con múltiples micrófonos [4]. A continuación se detallan las características generales del programa enumeradas en su manual de usuario.

- Corrige para distancias hasta de ~112 pies/ ~34 metros o un corrimiento de ± 100ms.
- El modo *Dynamic* permite correcciones continuas de fase/tiempo para hablantes o cámaras en movimiento.

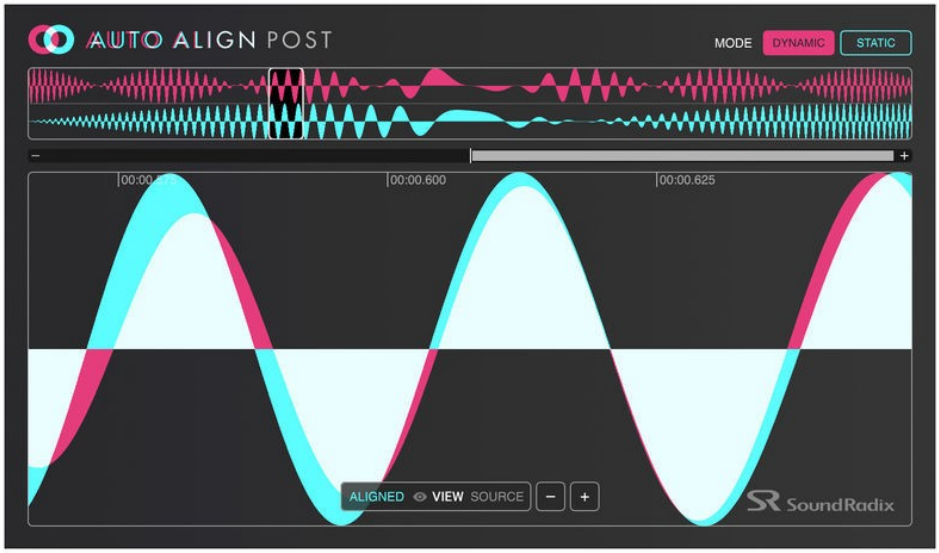

Figura 5. Captura de pantalla del programa Auto Align Post de Sound Radix.

3. MARCO TEÓRICO

En este capítulo se describe la teoría de los distintos procesos involucrados en la suma y alineación de señales de audio.

3.1. INTERFERENCIAS FRECUENCIALES

Al sumar dos señales de audio similares, la incorrecta alineación entre ambas puede generar interferencias [15]. Estas interferencias se manifiestan como atenuaciones y realces de frecuencias dentro del espectro audible.

3.1.1 Filtro Peine

El fenómeno audible más notorio que produce la suma de señales similares desplazadas en tiempo es el Filtro Peine. Este desplazamiento genera una primera cancelación ubicada en la frecuencia cuyo período sea el doble de la diferencia en tiempo entre pistas. Dicha frecuencia será la fundamental del filtro (f_0), y a continuación de f_0 aparecen cancelaciones en las frecuencias armónicas impares y realces en las pares [15].

En la Figura 6 se muestra la suma de dos señales de ruido blanco muestreadas a 44100 Hz. Ambas señales son idénticas, pero se ha introducido un corrimiento entre ellas de 1000 muestras (22 milisegundos), previamente a la suma. Como se puede ver, la primera cancelación ocurre alrededor de 22 Hz, que a la frecuencia de muestreo dada es de 2000 muestras por período.

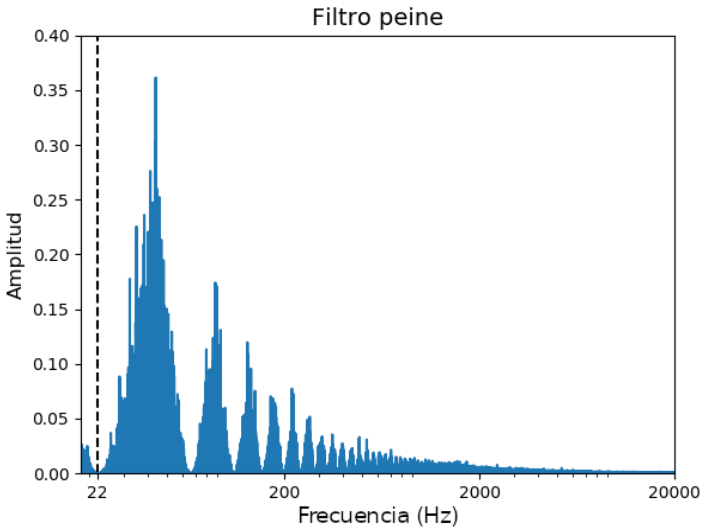

Figura 6. Filtro peine formado por la por la suma desfasada de una señal de ruido.

La frecuencia central del filtro peine depende de la diferencia en tiempo entre ambas señales. Si esa diferencia en tiempo varía, también lo hará la frecuencia central del filtro peine. Se puede ilustrar este fenómeno suponiendo una grabación de diálogo utilizando 2 micrófonos donde uno es fijo y el otro se desplaza a velocidad constante. Por más que ambas pistas de audio se encuentran inicialmente alineadas, la diferencia de tiempo generada a partir del movimiento de una de ellas genera variaciones en la frecuencia central del filtro peine proporcionales a la velocidad de desplazamiento del micrófono. Las Figuras 7 y 8 representan este fenómeno gráficamente para una velocidad de micrófono igual a 1 m/s en dos frecuencias diferentes. La señal azul representa la señal captada por el micrófono fijo y la roja la señal captada por el micrófono en movimiento.

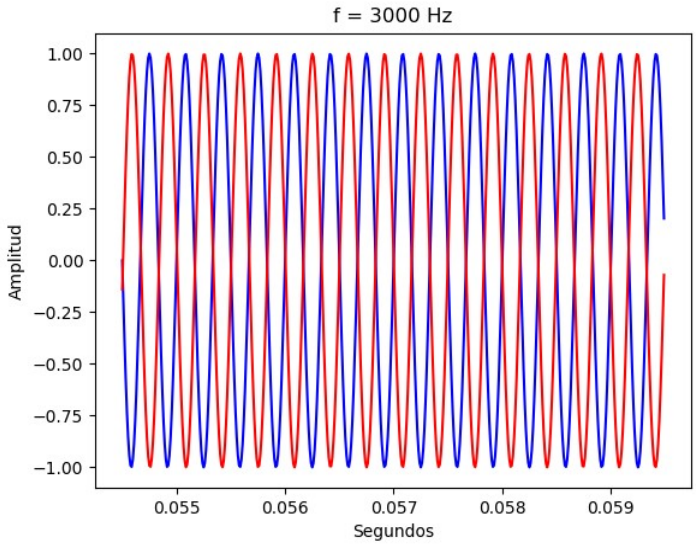

Figura 7. Corrimientos generados en un tono puro de 3000 Hz por un movimiento de fuente de 1 m/s.

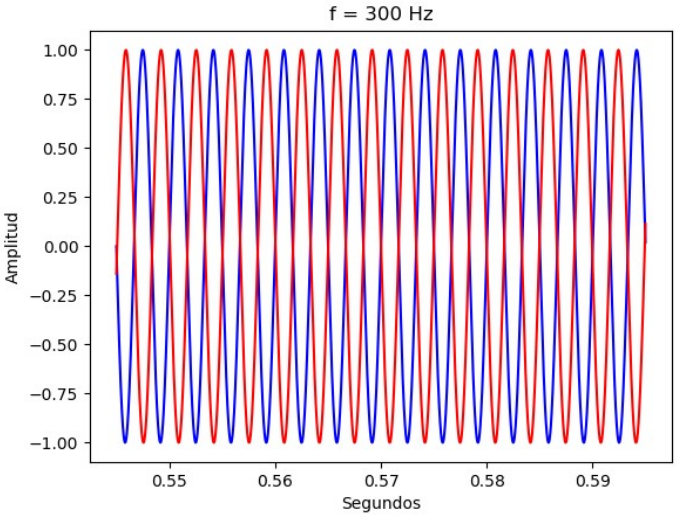

Figura 8. Corrimientos generados en un tono puro de 300 Hz por un movimiento de fuente de 1 m/s.

3.1.2. Relación de distancia variable

En una grabación de diálogo en la que se pretende registrar el sonido utilizando un micrófono de caña y uno de solapa, se da una situación particular en la que tanto la fuente como los dos micrófonos varían su posición. Esto genera que la distancia que existe entre cada micrófono y la fuente no solamente sea variable, sino que dichas variaciones no tienen a priori una correlación entre sí (Figura 9). La distancia entre el micrófono de caña y la fuente (DB) puede alterarse de forma independiente a la distancia entre el micrófono de solapa y la fuente (DL). Estos corrimientos generan filtros peine de frecuencia fundamental variable y comportamiento complejo al sumar las señales.

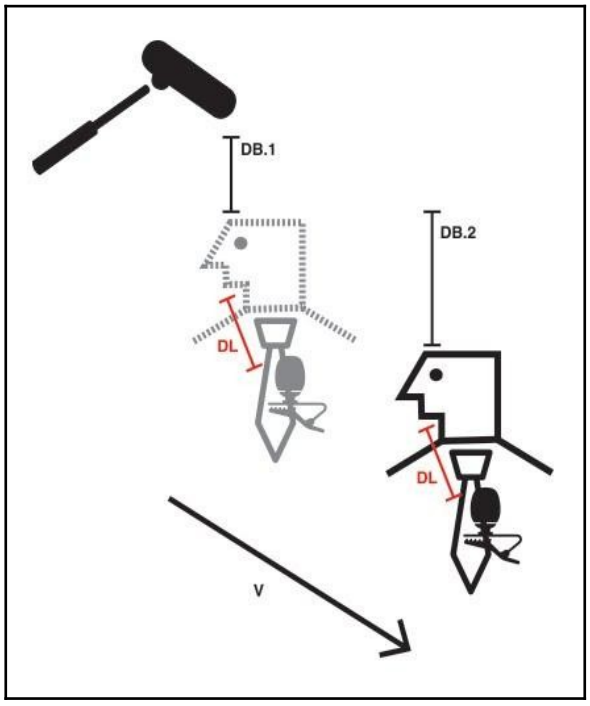

Figura 9. Variaciones de distancia entre el hablante y el micrófono de solapa y el de caña.

Al variar la distancia entre la fuente y el micrófono se genera un cambio en la frecuencia conocido como efecto Doppler. De acuerdo con [16] el oído humano percibe diferencias de frecuencia superiores al 0.5% de la frecuencia original. Teniendo en cuenta que este método de grabación de diálogo suele utilizarse en situaciones de bajo movimiento, se considera una velocidad máxima de 2 m/s. A esta velocidad las variaciones de frecuencia son menores al 0.5% dentro del espectro audible, por lo tanto

no son perceptibles. De todas formas, al sumar las diferentes señales, dichas variaciones son suficientes como para generar filtros peine.

3.2. HERRAMIENTAS DEL PROCESAMIENTO DE SEÑALES

En el diseño del algoritmo de alineación se utilizan muchas herramientas clásicas del procesamiento digital de señales como son la correlación cruzada, el filtrado de respuesta finita al impulso (FIR) y la normalización. A continuación se describen brevemente los aspectos fundamentales necesarios para esta tesis.

3.2.1. Correlación cruzada

En el estudio de señales, la correlación cruzada es la función matemática que permite evaluar la similitud entre dos señales. En la ecuación 1 se muestra la expresión de la correlación cruzada para dos señales discretas reales *f[n]* y *g[n]*, de longitud infinita.

$$R_{fg} = \sum_{m=-\infty}^{\infty} f[m]\,g[m+n] \tag{1}$$

En la aplicación a dos señales discretas finitas, la longitud de la señal de correlación será la suma de las respectivas longitudes de las señales *f* y *g*, menos dos muestras. El valor máximo del vector de correlación indica la posición de mayor similitud entre ambas señales.

La Figura 10 muestra la correlación entre dos señales iguales con un corrimiento de 1000 muestras entre ellas. El pico en la gráfica de correlación (rojo) se ve corrido del centro debido a esto. El desplazamiento entre pistas se obtiene calculando la diferencia entre la posición horizontal del pico y el centro de la señal.

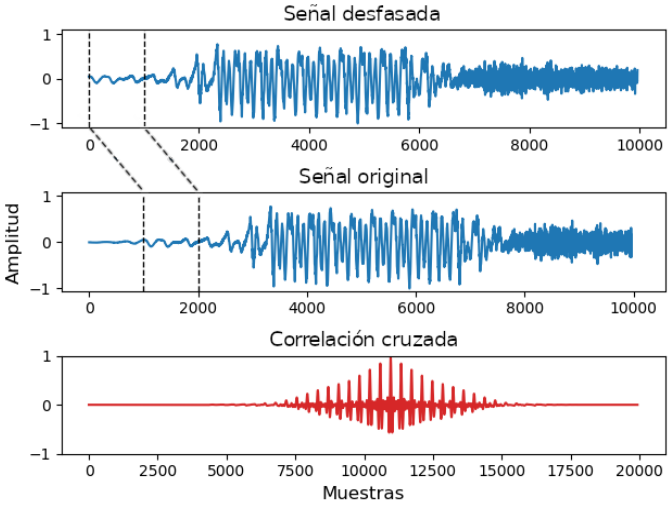

Figura 10. Señales desfasada, original y su correlación.

3.2.2. Filtros FIR

El filtrado constituye una herramienta fundamental del procesamiento digital de señales. A través de ella es posible eliminar las componentes de frecuencias irrelevantes para nuestro objetivo. En la presente tesis se utilizaron exclusivamente filtros digitales FIR debido a su capacidad para mantener constante el retardo de grupo (fase lineal) de la señal filtrada [17]. Los filtros FIR de fase lineal generalizada poseen una respuesta al impulso con una cantidad finita de términos. La Figura 11 muestra la estructura básica de un filtro FIR, y en la ecuación 2 muestra la función de transferencia de un filtro FIR.

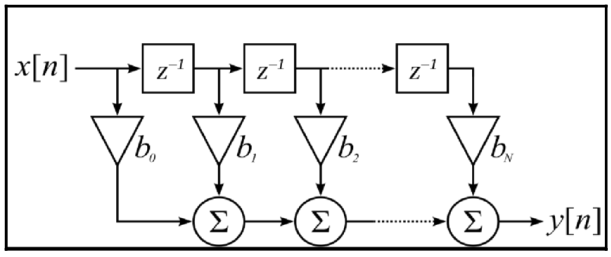

Figura 11. Estructura básica de los filtros FIR.

A partir del análisis de esta ecuación, mediante la transformada Z, se puede observar que todos los polos quedan ubicados en el origen, haciendo que estos filtros sean absolutamente estables. En cambio, los ceros, que dependen solamente de los coeficientes b_k de la ecuación, pueden ubicarse en cualquier lugar del plano Z.

$$y_n = b_0\,x(n) + b_1\,x(n-1) + \ldots + b_{M-1}\,x(n-M+1) = \sum_{k=0}^{M-1} h(k)\,x(n-k) \tag{2}$$

donde:

y_n es la señal filtrada

b_k son los coeficientes del filtro

$x(n)$ es la señal de entrada

n es la longitud de la señal de entrada

M es la longitud de la respuesta al impulso

3.2.3. Normalización RMS

Con el objetivo de generar un estimador de calidad de la correlación entre dos señales, se aplica una normalización de cada señal por su valor cuadrático medio. En el caso ideal donde ambas señales sean idénticas, la correlación entre ambas señales luego de ser normalizadas a su valor RMS, será igual a 1. Es válido considerar que mientras más cercano a 1 sea el valor de la correlación entre dos señales normalizadas RMS, mayor será el grado de la correlación.

$$y_{norm}[i] = \left[\frac{y_1}{\sqrt{y_1^2 + y_2^2 + \ldots + y_n^2}}; \frac{y_2}{\sqrt{y_1^2 + y_2^2 + \ldots + y_n^2}}; \ldots; \frac{y_n}{\sqrt{y_1^2 + y_2^2 + \ldots + y_n^2}} \right] \tag{3}$$

$$y_{cor} = \sum_{i=0}^{n} y_{norm}^2 \tag{4}$$

$$y_{cor} = \frac{y_1^2}{y_1^2 + y_2^2 + \ldots + y_n^2} + \frac{y_2^2}{y_1^2 + y_2^2 + \ldots + y_n^2} + \ldots + \frac{y_n^2}{y_1^2 + y_2^2 + \ldots + y_n^2} = 1 \tag{5}$$

Donde, y_{norm} es la señal normalizada, y_{cor} es el valor de la correlación, y_i es el valor de la muestra i y n es el número total de muestras.

4. DISEÑO DE ALGORITMO

A continuación se describen los diferentes procesos que conforman el algoritmo propuesto para la detección y posterior corrección de la desalineación.

4.1. ESTRUCTURA GENERAL

En esta sección se describe de manera general cada uno de los procesos que conforman el algoritmo planteado. En la Figura 12 se muestra su diagrama en bloques.

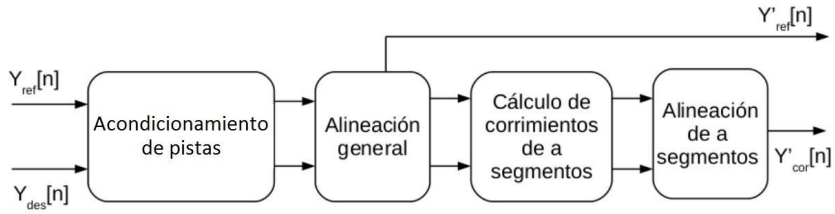

Figura 12: Diagrama en bloques general de algoritmo

El proceso comienza con las señales originales $Y_{ref}[n]$ e $Y_{des}[n]$, que arbitrariamente se definen como de referencia y desplazada respectivamente. Como salida se tiene $Y'_{ref}[n]$, que es la señal de referencia luego del proceso de alineación general e $Y'_{cor}[n]$ que es la señal corregida (señal desplazada luego del proceso de alineación).

En primer lugar, se hace un acondicionamiento de ambas señales con el objetivo de reducir las diferencias espectrales entre ellas y normalizar sus amplitudes. Tanto en la etapa de normalización como en la etapa del filtrado se utilizan los mismos parámetros para cada señal.

Finalizada la etapa de acondicionamiento, se realiza una alineación general, desplazando una de las dos señales para determinar un punto de inicio con corrimiento cero entre ellas. Esta posición funciona como referencia para la siguiente etapa de alineación en la que se calculan los corrimientos entre ventanas o segmentos de la totalidad de las señales. La Figura 13 muestra una comparación entre un segmento de las señales de referencia y desplazada antes y después de la alineación general. Las lineas punteadas forman una ventana dentro de la cual se considera que ambas señales se encuentran alineadas.

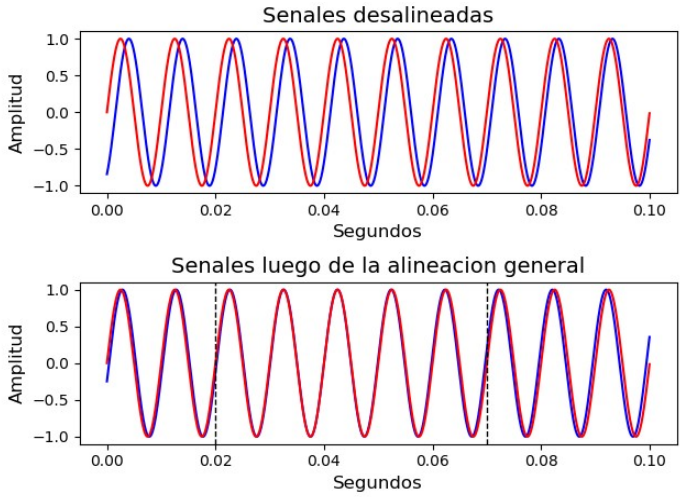

Figura 13. Comparación de señales antes y después del proceso de alineación general.

En la etapa de cálculo de corrimientos de a segmentos, se divide la señal de referencia en ventanas de igual cantidad de muestras; a esta cantidad de muestras fijas se la denomina "*lv*". Este proceso se ejemplifica gráficamente en la Figura 14, para un largo "*lv*" de 200 muestras.

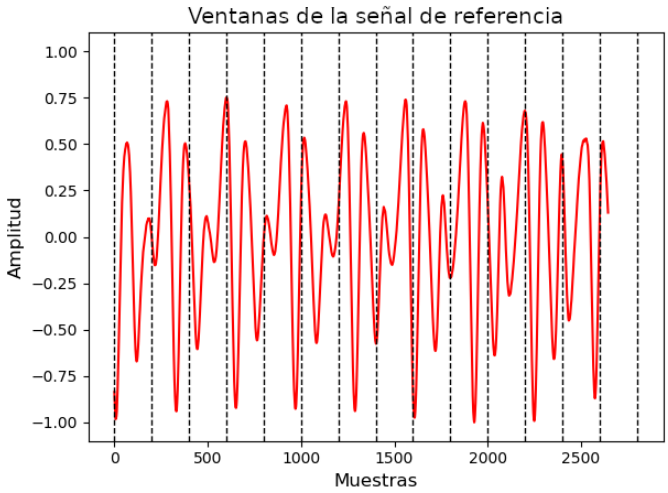

Figura 14. Señal de referencia segmentada en ventanas de 200 muestras.

Cada ventana de "*lv*" muestras de la señal de referencia se corresponde con una ventana de la señal desplazada de longitud *"lv + r[n]"*; Donde "*r[n]*" es el máximo corrimiento posible entre ventanas. Este valor puede variar entre diferentes pares de ventanas, dependiendo de los umbrales operativos descritos en la sección 4.4.3.

El máximo valor de correlación cruzada calculado entre ambas ventanas indica la posición de mayor similitud y, por lo tanto, su localización permite conocer el corrimiento.

Los corrimientos entre pares de segmentos de Y'_{ref} e Y'_{des} se calculan de forma sucesiva y en orden.

Por último, se realiza la alineación de segmentos de la señal desplazada Y_{des} de acuerdo con los corrimientos calculados. El algoritmo devuelve dos señales, la primera es la señal de referencia alineada por la primera etapa y la segunda es la señal corregida.

En la Figura 15 se muestran, a modo de ejemplo, las señales antes y después del proceso de alineación.

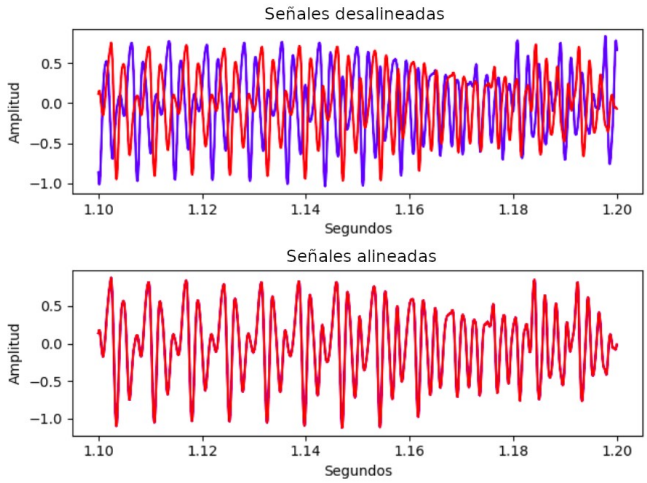

Figura 15. Señales de referencia y desplazada Vs señales de referencia y corregida.

A continuación se detalla el funcionamiento de cada una de las etapas y procesos involucrados.

4.2. ACONDICIONAMIENTO DE PISTAS DE AUDIO

En esta etapa, se describen en detalle los algoritmos que procesan los audio de entrada (Y_{ref} e Y_{des}) con el fin de reducir las diferencias espectrales existentes entre ellas y normalizar sus amplitudes. Ambas señales están acondicionadas de igual manera.

4.2.1. Normalizado

En primer lugar, se normalizan las señales de entrada para reducir las diferencias de amplitud entre ellas. Se calcula el cociente entre cada una y sus respectivos valores máximos (Ecuación 6); de esta forma la amplitud queda acotada entre -1 y 1.

$$Y_{norm}[n] = \frac{Y[n]}{max(Y)} \tag{6}$$

Donde Y_{norm} es la señal normalizada, Y es la señal original y $max(Y)$ es el máximo valor absoluto de la señal Y.

4.2.2. Filtrado

Con el objetivo de reducir las diferencias espectrales entre las señales y eliminar la información innecesaria presente en las tomas, se filtran las frecuencias irrelevantes para el propósito del algoritmo. Aunque las grabaciones corresponden al mismo evento sonoro, fueron registradas con micrófonos diferentes ubicados en distintas posiciones, atravesando funciones de transferencia independientes. Es por esto que su espectro puede presentar diferencias significativas.

Teniendo en cuenta que el objetivo es alinear pistas de diálogos, se consideran todas las frecuencias superiores o inferiores al espectro propio de la voz humana como innecesarias. Las frecuencias fundamentales de la voz se encuentran comprendidas entre 85 Hz y 180 Hz en hombres adultos y entre 165 Hz y 255 Hz para mujeres adultas [18]. Por su parte, las frecuencias armónicas de la voz pueden llegar hasta los 8 kHz [19]. A partir de esto, se aplica un filtro pasa bajo sintonizado a 8 kHz. La Figura 16 muestra un ejemplo de espectro de voz masculina con frecuencias desde 80 Hz hasta 5 KHz.

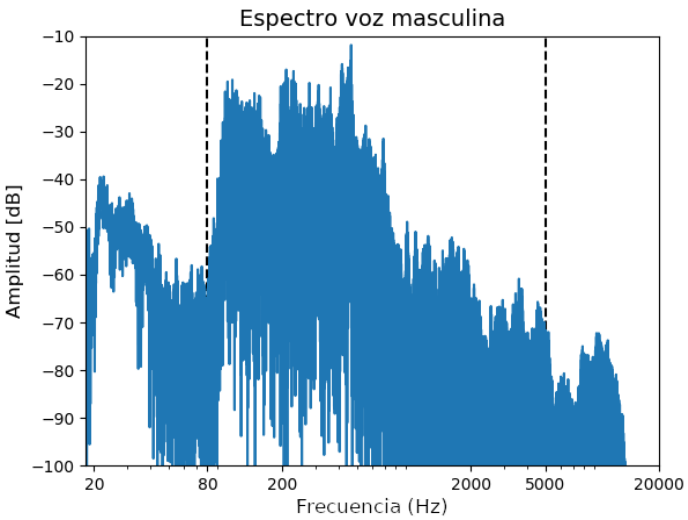

Figura 16. Espectro de la voz humana masculina

Se aplica también un filtro pasa alto, pero su frecuencia de corte depende de la longitud de la ventana elegida y de la frecuencia de muestreo de la señal. Se calcula la frecuencia cuyo período en muestras sea igual a la mitad de la longitud de la ventana

seleccionada (Ecuación 7). De esta forma, la mínima frecuencia puede cumplir dos ciclos enteros dentro de una ventana de correlación.

$$f_{ci} = \frac{f_s}{lv * 2} \tag{7}$$

Donde f_s es la frecuencia de muestreo de las señales y f_{ci} es la frecuencia de corte del filtro pasa alto.

Para realizar el filtrado se aplican filtros de respuesta al impulso finita (FIR) de fase lineal [20], utilizando el método de diseño por ventanas. Para ambos casos se aplica una ventana *Blackman* en la generación de la respuesta al impulso de los filtros, por su desempeño superior para aplicaciones en audio digital [17]. La implementación se realizó mediante la función *signal.firwin* de la biblioteca *scipy* de *Python* [21].

4.3. ALINEACIÓN GENERAL

Para la correcta identificación de corrimientos entre ventanas, es necesario partir de una posición donde las dos pistas estén alineadas. Se busca entre las señales de referencia y desplazada, el par de ventanas de longitud "lv" con el máximo valor de correlación. Luego se alinean ambas pistas para que el corrimiento entre el par hallado sea cero; y el resto de los corrimientos se calculan a partir de esta posición inicial.

Esto se logra calculando la correlación entre dos ventanas de las señales Y'_{ref} e Y'_{des} de forma cíclica, reduciendo en cada ciclo el tamaño de las ventanas a la mitad de su longitud hasta un valor mínimo de "*lv*". Cada ventana se normaliza utilizando su valor cuadrático medio (RMS), de forma que el valor de correlación calculado sea independiente de la amplitud de la señal en el par de ventanas a correlacionar. En cada ciclo se alinean ambas pistas a su posición de máxima correlación. De todos los calculados, el segmento de mayor correlación ($n_{inicial}$), es utilizado para la alineación global de las señales $Y'_{ref}[n_{inicial}]$ y $Y'_{des}[n_{inicial}]$.

En la Figura 17 se muestra un ejemplo gráfico del proceso de alineación global. La Figura representa un solo ciclo del algoritmo usado para obtener $n_{inicial}$. Las señales de correlación (verde y violeta) representan las señales de correlación entre la señal de referencia (azul) y las mitades de la señal desplazada (amarillo y rojo). Como el valor de la

señal verde es mayor que el de la señal violeta, para el próximo ciclo se repetirá el proceso utilizando solamente la primera mitad de la señal desplazada.

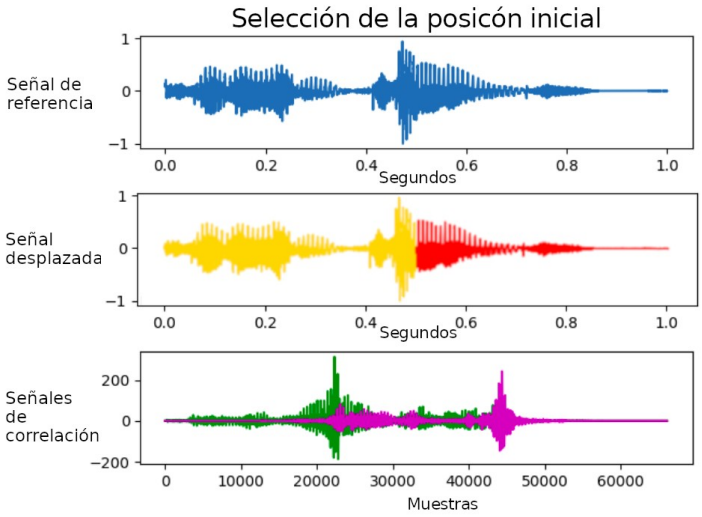

Figura 17. Ejemplo gráfico de alineación global.

4.4. CÁLCULO DE CORRIMIENTOS DE A SEGMENTOS

Una vez realizada la alineación global y definida la primera ventana, el algoritmo calcula los corrimientos de forma sucesiva, primero de las ventanas posteriores y luego de las anteriores a $n_{inicial}$. El resultado es un vector (t_a) con los valores de corrimientos entre cada par de ventanas de las señales $Y'_{des}[n]$ e $Y'_{ref}[n]$. La Figura 18 muestra los diferentes procesos involucrados en esta etapa.

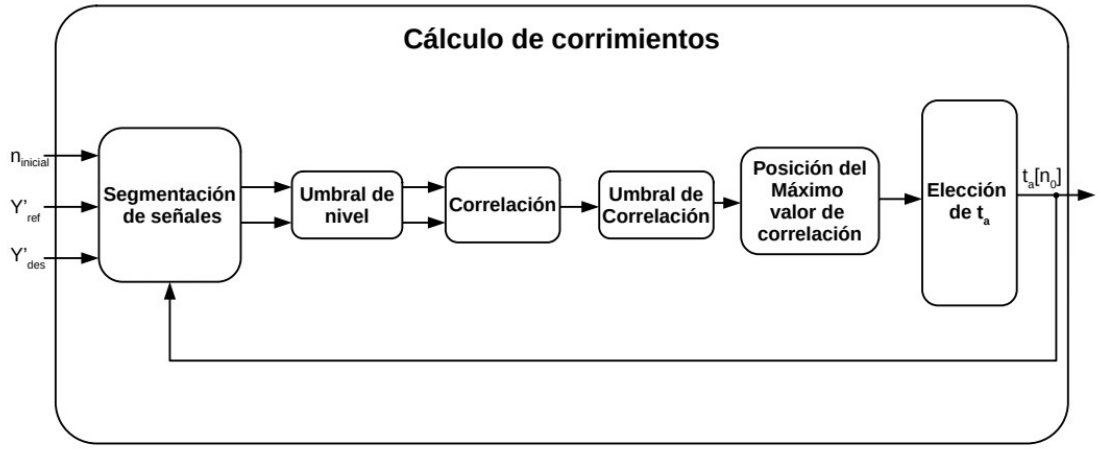

Figura 18. Diagrama de bloques del cálculo de corrimientos

A continuación, se describen los procesos involucrados en la obtención del vector de corrimientos a partir de las señales Y'_{ref} e Y'_{des}.

4.4.1 Elección de "*lv*"

Para la correcta elección de la longitud de ventana a correlacionar, es necesario considerar la aplicación particular del algoritmo propuesto. Si se toman ventanas de pocas muestras, la correlación no tendría información suficiente como para detectar correctamente la posición de mayor similitud entre señales. Por otra parte, si se toman ventanas muy grandes, es posible que los corrimientos entre las señales producto del movimiento de la fuente y los micrófonos generen inversiones de fase dentro de una misma ventana, generando efectos audibles (principalmente en las frecuencias más altas). La Figura 19 muestra las cancelaciones generadas dentro de una ventana a partir de una elección incorrecta de "*lv*". En esa Figura se puede observar que por más que ambas señales se encuentren alineadas en el centro de la ventana, los corrimientos acumulados en los extremos pueden generar cancelaciones audibles.

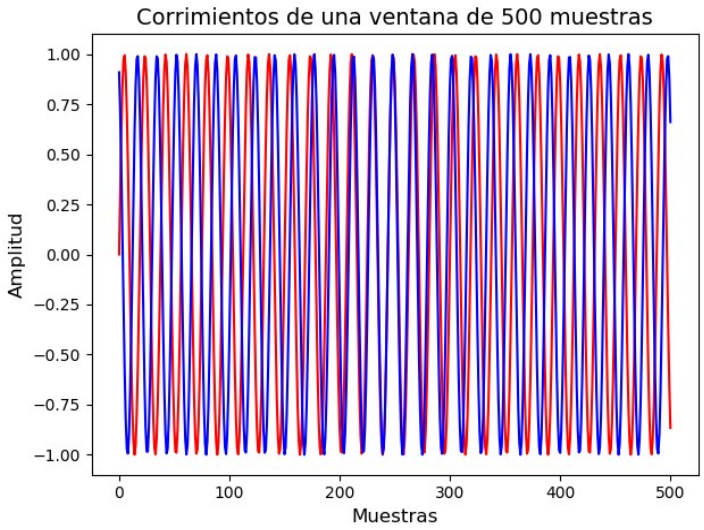

Figura 19. Corrimientos dentro de una ventana de 500 muestras

Como se vió previamente, la voz humana carece de información significativa para frecuencias superiores a 8 kHz. Por lo tanto, se busca evitar corrimientos significativos para frecuencias inferiores dentro de la longitud "*lv*". Se elige como criterio un corrimiento máximo permitido de un cuarto de ciclo. Un cuarto de ciclo de 8 KHz requiere

31.25 µs para desarrollarse. Por ello, el corrimiento máximo posible dentro de una ventana debe ser menor a este valor.

Se considera también que para esta aplicación, no es necesario tener en cuenta corrimientos generados por movimientos mayores a 4 m/s. Tomando la velocidad del sonido como 343 m/s, se calcula el corrimiento por segundo producto de dicho movimiento de la siguiente manera:

$$\frac{v_f . 1\,seg}{C} = t_0 \tag{8}$$

Donde, v_f es la velocidad de la fuente, en este caso 4 m/s, C es la velocidad del sonido, t_0 es la diferencia en tiempo de arribo entre una señal fija y una señal con un movimiento a velocidad constante v_f. El corrimiento temporal generado por este movimiento resulta en 11,7 ms por cada segundo transcurrido. Se usa la siguiente expresión para calcular la máxima longitud de la señal de correlación:

$$\frac{2.t_c . fs}{t_0} = n \tag{9}$$

Donde, t_c es igual a un cuarto del período de la frecuencia máxima de interés, f_s es la frecuencia de muestreo de las señales a procesar, t_0 es la diferencia en tiempo de arribo entre una señal fija y una señal con un movimiento a velocidad constante de 4 m/s, n es el máximo número de muestras que pueden tener las señales a correlacionar para evitar cancelaciones en frecuencias inferiores a 8 kHz. La ecuación 9 está multiplicada por 2 debido a que la alineación se realiza con el centro de la ventana y no desde los extremos. De esta ecuación se desprende que el tamaño máximo en muestras de la ventana temporal para señales cuya frecuencia de muestreo es de 44.1 kHz, es de 235 muestras. Partiendo de estos valores, se elige un valor de ventana temporal igual a 200 muestras.

4.4.2. Máximo corrimiento entre ventanas consecutivas

Dada la aplicación particular del algoritmo, se pueden establecer valores de máximos corrimientos permitidos entre cada par de ventanas consecutivas. De esta forma se reduce significativamente el tiempo de procesamiento y se evita el cálculo innecesario de valores de correlación cruzada. El máximo corrimiento entre ventanas (*mcv*) depende

de la longitud de ventana "*lv*" y de la velocidad de los movimientos de los micrófonos y la fuente durante la grabación del diálogo. Se utiliza el mismo criterio que en 4.4.1, donde la velocidad máxima considerada es de 4 m/s. Así, se puede obtener el corrimiento máximo en muestras posible por ventana con la Ecuación 10.

$$mcv = \frac{lv \,.4\, m/s}{C} \tag{10}$$

Donde, *C* es la velocidad del sonido en el aire.

Con una longitud de ventana "*lv*" igual a 200 muestras el máximo corrimiento por ventana resulta igual a 2,3 muestras. En otras palabras, el corrimiento entre dos pares de ventanas consecutivas bajo condiciones normales no debe exceder este valor. De todas formas, existen dos situaciones especiales donde se aceptan valores de corrimiento entre ventanas mayores al "*mcv*" calculado. Ambas situaciones dependen de los umbrales operativos descritos a continuación

4.4.3. Umbrales operativos

Se establecen dos condiciones que cada par de ventanas debe cumplir para considerar válido el valor de corrimiento calculado. Cuando un par no cumple alguna de las condiciones, se iguala su valor de corrimiento al último calculado en condiciones normales.

La primera condición funciona como una compuerta de ruido, omitiendo los corrimientos calculados para segmentos silenciosos. El umbral preestablecido se compara con el cociente entre el valor RMS de la ventana procesada y el valor RMS de la totalidad de la señal. El umbral es el mismo, tanto para la señal de referencia como para la señal desplazada. Si el cociente descrito para cualquiera de las dos señales resulta menor que el umbral establecido, no se realizará el cálculo de correlación y el corrimiento relativo para esa ventana será igual a cero.

La segunda condición es un umbral de valor mínimo de correlación. Dado que las señales a correlacionar fueron previamente normalizadas a sus respectivos valores RMS, el valor máximo de correlación posible es 1. En caso de tener valores de correlación

cercanos a cero, se puede considerar que la coherencia del cálculo es insuficiente. A partir de esto, se establece un umbral de correlación mínima.

En ambos casos se suma uno al valor del máximo corrimiento entre ventanas (mcv) para el cálculo de correlación de la ventana siguiente.

$$r\left[n_0\right]=mcv+c_g\left[n_0\right]+c_c\left[n_o\right] \tag{11}$$

Donde $r[n_0]$ es el máximo corrimiento para la ventana n_0, $c_g[n_0]$ es el coeficiente por umbral de ruido para la ventana n_0 y $c_c[n_0]$ el coeficiente por umbral de correlación para la ventana n_0.

El aumento en el máximo corrimiento entre ventanas se debe a que, de no ser capaz de detectar el corrimiento entre un par de ventanas, el par siguiente puede estar arrastrando un corrimiento no calculado. Una vez que se llega a un par de ventanas que superan ambos umbrales establecidos, los coeficientes vuelven a ser cero.

La correcta elección de ambos umbrales es determinante para identificar correctamente los corrimientos entre señales. Dependiendo de las características dinámicas y espectrales de las señales de entrada, es posible que los valores de umbrales que optimicen la identificación de corrimientos varíe. Por este motivo, el algoritmo realiza múltiples repeticiones evaluando diferentes combinaciones de valores de umbrales. Se eligen 5 valores de cada umbral generando un total de 25 combinaciones posibles. Se utiliza el valor de correlación entre la señal de referencia y la corregida para determinar cuál combinación funciona mejor en cada caso. Es importante destacar que este proceso genera un aumento considerable del tiempo de procesamiento del algoritmo debido a las múltiples repeticiones.

4.4.4. Segmentación de señales

En primer lugar, se segmenta la señal de referencia. La cantidad de ventanas generadas se calcula en función de la cantidad de muestras y de ”*lv*” de acuerdo con la Ecuación 12.

$$N=\frac{len\left(Y_{ref}\right)}{lv} \tag{12}$$

Donde *N* es el número total de segmentos y *len*(Y_{ref}) es la cantidad total de muestras de la señal de referencia. La cantidad de ventanas determina la cantidad de cálculos de corrimientos que se realizarán. La señal de referencia es separada en ventanas de muestras consecutivas de longitud "*lv*". Sus posiciones inicial y final se obtienen a partir de las siguientes ecuaciones (Ec. 13 y 14):

$$m_{iref} = n_0 . lv \tag{13}$$

$$m_{sref} = lv . (n_0 + 1) \tag{14}$$

Donde m_{iref} y m_{sref} son los números de muestras inferior y superior de la ventana respectivamente y n_0 es el número de ventana.

Las ventanas de las señales desplazadas se calculan de forma sucesiva, luego de obtener el valor de corrimiento de la ventana anterior. Esto se debe a que las ventanas de la señal desplazada toman en cuenta los corrimientos acumulados hasta el par de ventanas anterior. La posición inicial y final de cada ventana n_0 de la señal desplazada se obtienen a partir de las siguientes ecuaciones (Ec. 15 y 15).

$$m_{ides} = lv . n_0 - \frac{r[n_0]}{2} + t_a[n_0] \tag{15}$$

$$m_{fdes} = lv . (n_0 + 1) + \frac{r[n_0]}{2} + t_a[n_0] \tag{16}$$

Donde m_{ides} y m_{fdes} son las posiciones iniciales y finales de la ventana n_0 de la señal desplazada, $t_a[n_0]$ son los corrimientos acumulados hasta la ventana n_0 y $r[n_0]$ es el máximo corrimiento posible de la ventana n_0. La Figura 20 muestra la selección de las ventanas en cada señal.

El máximo valor de correlación obtenido entre ambas ventanas corresponde a la posición de mayor similitud entre ellas.

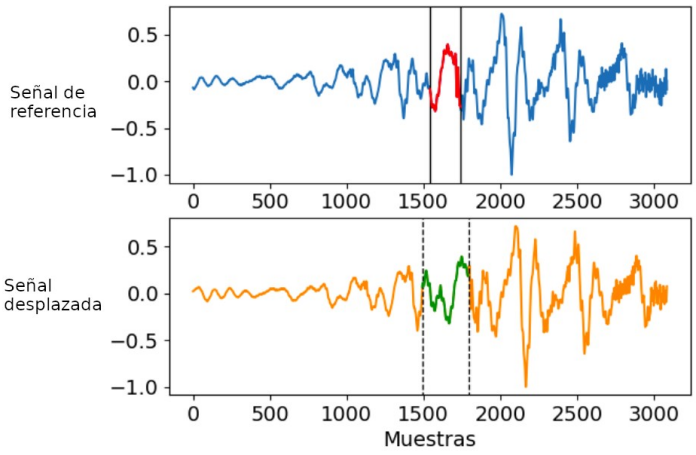

Figura 20. Selección de ventanas a correlacionar

4.4.5. Correlación cruzada

Se procede a realizar los cálculos de correlación cruzada entre ventanas de acuerdo a lo descrito en los capítulos anteriores. En cada repetición del ciclo se calcula la correlación entre una ventana de la señal de referencia y su correspondiente ventana de la señal desplazada. Esto genera un vector de valores de correlación por cada par de ventanas. La posición del máximo valor de correlación indica el valor del corrimiento relativo de la ventana n_0. La Figura 21 muestra la correlación entre un par de ventanas de las señales de referencia y desplazada. Las líneas punteadas marcan la posición de máxima correlación. En este ejemplo, la máxima correlación se da en la posición central entre ambas ventanas.

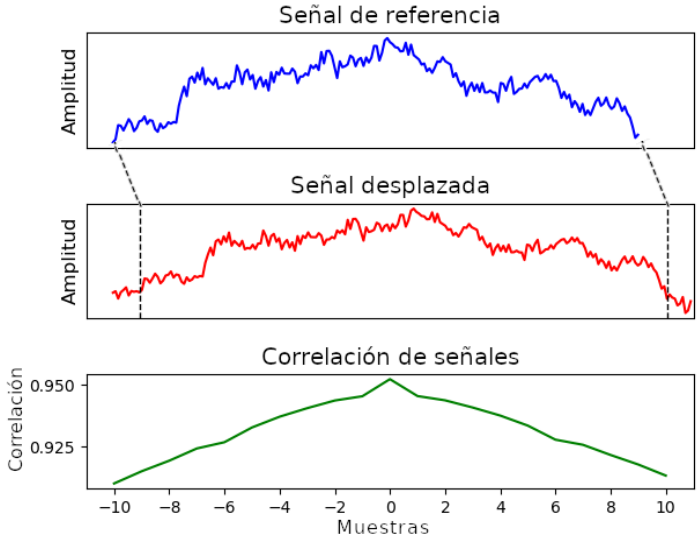

Figura 21. Correlación entre un par de ventanas.

Al sumar los corrimientos relativos anteriores, se obtienen los corrimientos absolutos por ventana, que son almacenados en el vector t_a para luego ser usados en la alineación de la señal desplazada. A continuación, se detallan los diferentes procesos involucrados en el cálculo de corrimientos entre señales.

Para cada valor de n_0 se genera un vector $cor[n_0]$ con los valores obtenidos de la correlación, que se calcula con la ecuación 17.

$$cor\left[n_0, i_0\right] = \sum_{j=0}^{lv} Y'_{ref}\left[n_0, j\right] \cdot Y'_{des}\left[n_0, r, j\right] \quad (17)$$

Donde *j* corresponde al número de muestra dentro de cada segmento, y los demás términos ya fueron definidos anteriormente. Una vez obtenido el vector cor[n_0] se busca la posición del máximo valor, que indica el número de muestras del desplazamiento entre los segmentos $Y_{ref}[n_0]$ y $Y_{des}[n_0,r]$.

La implementación de la correlación cruzada se hizo utilizando la función *correlate*, de la biblioteca *Numpy* [22], dentro del lenguaje *Python*.

4.5. ALINEACIÓN DE A SEGMENTOS

Para la alineación temporal, se utiliza una función que toma a la entrada la señal desplazada y los valores de corrimiento "t_a" para cada fragmento de longitud "lv". A partir de esto, se genera una nueva señal de audio formada por los fragmentos desplazados de la señal original. Dicha función opera de forma diferente si los corrimientos son positivos (señal desplazada hacia la derecha) o negativos (señal desplazada hacia la izquierda). Con el objetivo de evitar ruidos en los empalmes entre segmentos se agrega un *crossfade* en la unión entre cada ventana [23].

5. METODOLOGÍA DE EVALUACIÓN

Para evaluar la eficacia del algoritmo presentado se utilizan tres tipos de señales de audio diferentes: señales con corrimientos temporales incorporados digitalmente (Señal A), señales con corrimientos temporales generados a partir de desplazamiento en ambiente controlado (Señal B), y señales registradas durante grabaciones de diálogos para producciones audiovisuales reales (Señal C). La habilidad del algoritmo para alinear las pistas se evalúa de forma diferente para cada tipo de señal.

5.1. SEÑALES A

Se diseña un programa que introduce corrimientos temporales, en una señal de una voz masculina pregrabada [25], simulando la captación del sonido reproducido por una fuente móvil a velocidad constante "*v*". De esta forma, es posible conocer a priori el corrimiento exacto entre la señal original y la señal con corrimiento. El algoritmo no modifica ningún otro parámetro de la señal original.

El proceso consiste en estirar o comprimir temporalmente la señal, simulando las diferencias en tiempos de arribo que se darían a partir del movimiento de una fuente con respecto a un micrófono, lo que modifica el espectro de la señal original. Esa modificación es proporcional a la velocidad de la fuente; si la velocidad es baja, la alteración del espectro será imperceptible, pero al sumarla con la señal original, aparecerán filtros peine de frecuencia central variable similar a los descritos en la sección **3.1.1**.

En la Figura 22 se ven los gráficos de espectro de la suma entre una señal aleatoria antes y después de ser procesada. Los dos gráficos corresponden a momentos diferentes de la señal. Se puede observar como la frecuencia del filtro peine es diferente para cada momento.

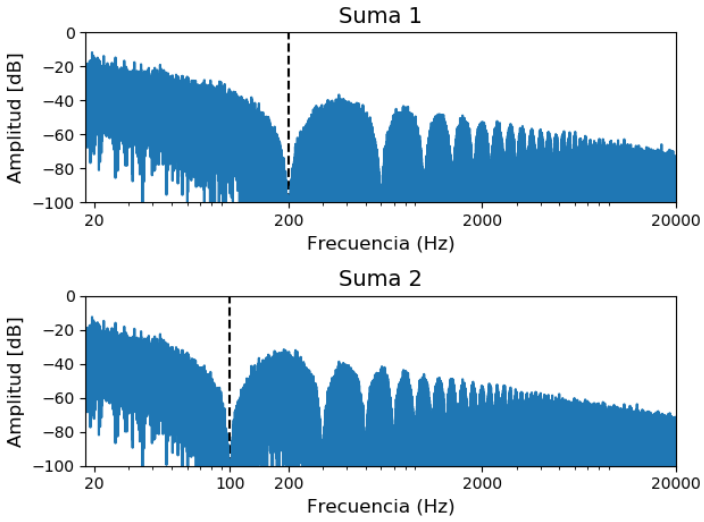

Figura 22. Espectros de diferentes fragmentos de la suma de una señal de ruido, antes y después de ser procesada.

Se utilizó una función que permite modificar la frecuencia de muestreo de la señal de entrada. Al reproducir la señal usando su frecuencia de muestreo original, se obtiene un corrimiento en frecuencias similar al que se obtiene con una fuente en movimiento y un receptor fijo. La siguiente expresión, ecuación 18, permite calcular la frecuencia de muestreo necesaria para simular el movimiento de una fuente a una velocidad constante "*v*".

$$f_{sf} = f_{si}\left(1 + \frac{v}{c}\right) \tag{18}$$

Donde f_{sf} y f_{si} son las frecuencias de muestreo final e inicial respectivamente, *v* es la velocidad elegida de desplazamiento de la fuente y *c* la velocidad del sonido en el aire.

Como señal de entrada se utilizó una grabación de voz masculina [24], y se generaron 8 señales de audio diferentes simulando variaciones de velocidad, sentido del movimiento y atenuación. La Tabla 1 muestra las características individuales de los archivos generados.

Tabla 1. Características de las señales del tipo A

Señal original	Número	Velocidad [m/s]	Atenuación (90%)
Grabación voz masculina	1	0.686	No
	2	0.686	Si
	3	1.029	No
	4	1.029	Si
	5	-0.686	No
	6	-0.686	Si
	7	-1.029	No
	8	-1.029	Si

Las velocidades elegidas (0.686 m/s y 1.029 m/s) corresponden a corrimientos de 88.2 y 132.3 muestras respectivamente, por cada segundo de audio digital muestreado a 44100 Hz. Se simularon los casos donde la fuente se acerca al micrófono (velocidad positiva) y donde la fuente se aleja (velocidad negativa). La atenuación se incorpora de forma lineal a lo largo de la señal hasta llegar al 90% de su valor original.

5.2. SEÑALES B

Las señales con corrimiento temporal generado a partir de desplazamiento en ambiente controlado, se grabaron en el estudio de grabación y mezcla Viet Music House [26]. Para la grabación se utilizó el micrófono Sennheiser Mkh60 de caña como micrófono fijo y el micrófono Oktava mk 012 en la mano de la persona hablante.

Se realizan siete grabaciones introduciendo un tipo de movimiento diferente para cada una. La Tabla 2 muestra las características de cada grabación.

Para las señales de movimiento lineal se incluyeron dos tipos de variaciones:

- Velocidad de caminata alta (∼1 m/s) y baja (∼0.6 m/s).
- Dirección de caminata (positiva y negativa).

Se registran también dos señales de voz manteniendo fija la fuente y variando la posición del micrófono de mano de forma ascendente y descendente. Por último, se registra el movimiento rotacional de la fuente hacia la izquierda y la derecha.

Todos los audios finales fueron registrados a una frecuencia de muestreo de 44.1 kHz y 32 bit de profundidad.

Tabla 2. Señales grabadas en ambiente controlado.

N.° de grabación	Movimiento	Velocidad	Dirección
1	Lineal	Alta	Negativa
2			Positiva
3		Baja	Negativa
4			Positiva
5	Micrófono	-	-
6	Rotacional	-	Izquierda
7		-	Derecha

5.3. SEÑALES C

Las señales de grabaciones de diálogos para producciones audiovisuales reales fueron provistas por el estudio de Foley y posproducción de audio para cine Ñandú sonido [27]. Corresponden a una grabación de un monólogo registrado con un micrófono de caña Sennheiser Mkh60 y un micrófono de solapa Sanken COS-11D, iguales a los descritos en **2.1.1**. La grabación corresponde a un monólogo de un hombre sentado en una silla; y los movimientos de la persona hablante y de los micrófonos son significativamente menores que para los dos grupos de señales anteriores.

5.4. MÉTODO DE EVALUACIÓN

Con el objetivo de evaluar la eficacia del algoritmo propuesto se emplean múltiples metodologías. En primer lugar, se calcula la correlación entre los pares de señales antes y después de realizar la alineación de pistas. Al alinear las señales, la similitud entre señales debería aumentar y, por lo tanto, su valor de correlación debería ser mayor. A su vez, se calcula la correlación entre cada par alineado utilizando el software comercial Auto-align Post (AAP), que se utiliza como punto de comparación.

También se realizan espectrogramas de la mezcla entre pares, antes y después de la alineación. El objetivo es observar en los gráficos la reducción de los filtros peine presentes antes de la alineación.

5.5. SOFTWARE IMPLEMENTADO

En el siguiente link se encuentran almacenados los algoritmos descritos. El archivo “tesisDisp1.0.py” ejecuta en orden las ecuaciones cargadas en “tesisFun4_4.py”. El

archivo “dopplersim2.py” contiene el algoritmo descrito en 5.1. https://github.com/FranMe/Tesis/

6. RESULTADOS Y CONCLUSIONES

A continuación se presentan los resultados para cada tipo de señales descritas en el capítulo 5.

6.1. SEÑALES A

La Figura 23 muestra los corrimientos calculados para el par de señales A1. Se puede observar que los corrimientos calculados siguen una tendencia lineal. Esto corresponde a la velocidad constante utilizada en la simulación de corrimientos. La pendiente de la recta indica que por segundo (220.5 ventanas de 200 muestras cada una) existe un desplazamiento de 88.2 muestras. Esto corresponde exactamente con el corrimiento esperado. Esto mismo se repite con las diferentes velocidades y atenuaciones de las señales del grupo A. A partir de esto, se puede concluir que el algoritmo propuesto es capaz de detectar los corrimientos generados digitalmente.

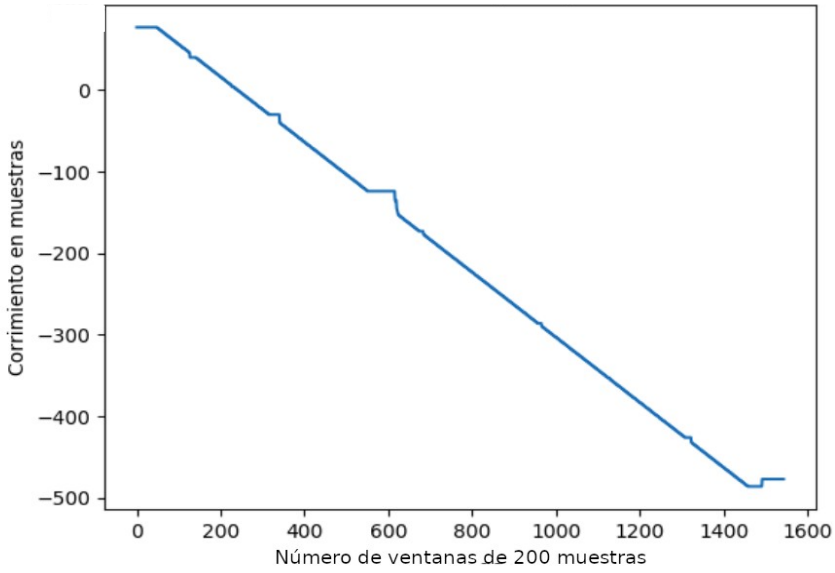

Figura 23. Corrimientos calculados entre la señal de referencia y desplazada A1.

Los cambios de pendiente (escalones) que se observan en la Figura 23, corresponden a ventanas con valores de amplitud o correlación inferiores a los umbrales operativos descritos en 4.3.3. Los cálculos de corrimientos entre pistas no son tenidos en cuenta si los segmentos a correlacionar son silenciosos o si el valor de la correlación es inferior a un umbral preestablecido. Al calcular los corrimientos de las diferentes ventanas de las señales tipo A, los valores de correlación se encuentran siempre por encima del

umbral. Los cambios de pendiente visibles en la Figura 23 corresponden a segmentos silenciosos.

La Tabla 3 muestra los diferentes cálculos de correlación de las señales A de referencia, desplazadas y corregidas. Previamente a la correlación, se realizó la normalización RMS; por lo tanto, el máximo valor posible es 1. La columna “Desplazada” muestra la correlación entre las señales de referencia y las señales desplazadas antes del proceso de alineación. La columna “Corregida” muestra la correlación entre la señales de referencia y la corregida por el algoritmo propuesto.

Tabla 3. Valores de correlación calculados a partir de las señales A de referencia, desplazadas, corregidas y umbrales optimizados.

Señal	N°	Descripción	Desplazada	Corregida	Umbral de amplitud	Umbral de correlación
A	1	Velocidad 0.686 m/s. Sin atenuación	0.131	0.995	0.02	0.08
	2	Velocidad 0.686 m/s. Con atenuación	0.142	0.985	0.02	0.08
	3	Velocidad 1.029 m/s. Sin atenuación	-0.095	0.984	0.05	0.11
	4	Velocidad 1.029 m/s. Con atenuación	-0.113	0.976	0.05	0.11
	5	Velocidad -0.686 m/s. Sin atenuación	0.133	0.993	0.02	0.06
	6	Velocidad -0.686 m/s. Con atenuación	0.143	0.982	0.02	0.06
	7	Velocidad -1.029 m/s. Sin atenuación	-0.095	0.985	0.02	0.12
	8	Velocidad -1.029 m/s. Con atenuación	-0.113	0.974	0.02	0.12

En todos los casos, el valor de correlación aumentó considerablemente luego de realizar el proceso de alineación con el algoritmo propuesto, llegando a valores cercanos al máximo posible para la situación propuesta.

En primer lugar, se observa que los valores de correlación entre las señales de referencia y las corregidas disminuyen con el aumento de la velocidad. Esto puede

deberse a la deformación temporal dentro de cada ventana de 200 muestras a correlacionar. Al ser mayor la velocidad, aumenta la deformación. Con respecto a las señales atenuadas, la situación es similar. Las señales que fueron atenuadas, además de desplazadas, presentan valores de correlación un poco más bajos que las señales sin atenuar. Nuevamente, esta diferencia puede deberse a las diferencias que la atenuación genera dentro de cada ventana a correlacionar, y no con una falla en la detección del corrimiento. Además de lo planteado, las diferencias entre el valor de correlación ideal y el obtenido tienen que ver con los segmentos donde los parámetros de amplitud y/o correlación se encuentran por debajo de los umbrales establecidos.

Las columnas “Umbral de amplitud” y “Umbral de correlación” muestran los valores optimizados de los umbrales para cada par de señales.

La Figura 24 muestra los espectrogramas de la mezcla de las señales de referencia y desplazada A1, antes y después de la alineación usando el algoritmo propuesto. En la figura de la izquierda se puede observar el filtro peine de frecuencia central variable. Las líneas más oscuras son las zonas de cancelaciones. En la figura de la derecha esas líneas desaparecieron casi en su totalidad. Esta misma tendencia se mantiene para todas las combinaciones de velocidades y atenuaciones de las señales A.

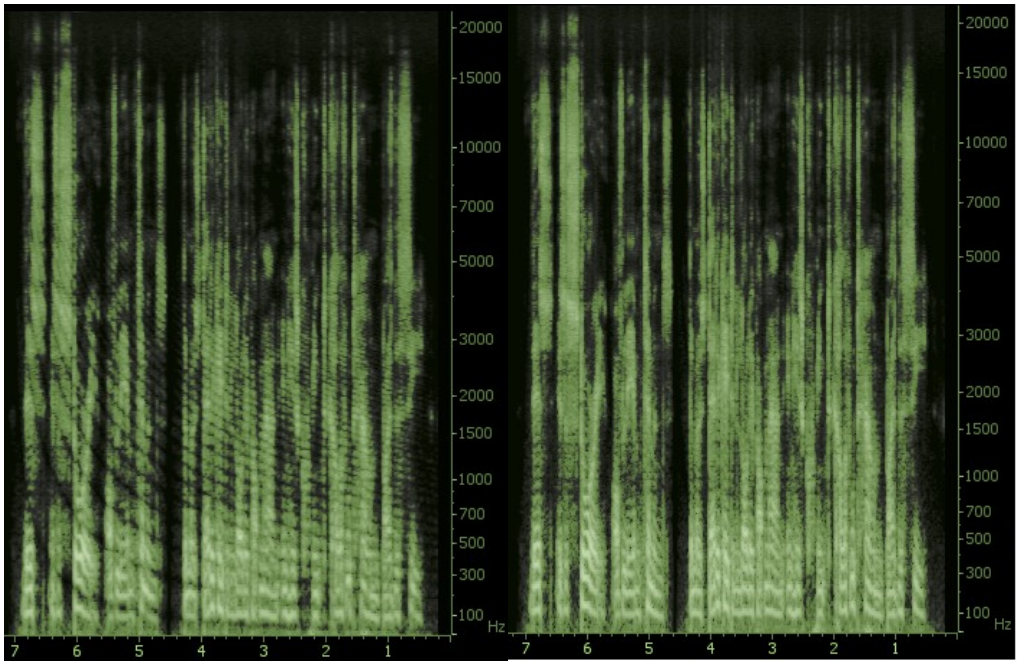

Figura 24. Comparación de espectrogramas.

Si se comparan las señales antes y después del procesamiento, la tendencia se mantiene. Las Figuras 25 y 26 muestran esta comparación en el caso donde hay una señal de voz presente y en el caso silencioso donde el valor RMS de las ventanas no supera el umbral preestablecido. Se puede ver que en la primera, las señales de referencia y corregida se superponen casi por completo. Por otro lado, en el segmento silencioso existe una diferencia más notoria entre la señal de referencia y la corregida.

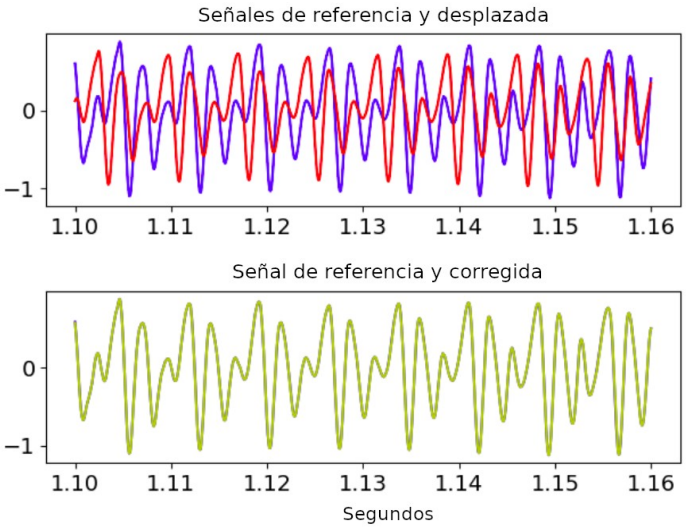

Figura 25. Comparación de las señales originales y corregidas en un segmento con voz.

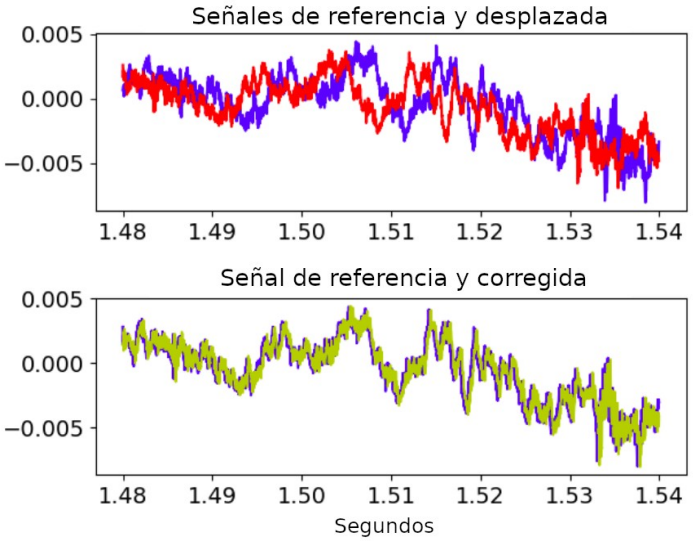

Figura 26. Comparación de las señales originales y corregidas en un segmento silencioso.

A partir de lo anterior, se concluye que el algoritmo propuesto fue eficaz para corregir los corrimientos introducidos digitalmente para todos los casos propuestos.

6.2. SEÑALES B

La Figura 27 muestra los corrimientos calculados por el algoritmo entre las señales de referencia y desplazada B1. En este caso los corrimientos introducidos se deben a un movimiento lineal en dirección negativa (opuesta al micrófono fijo) a velocidad aproximadamente constante. Al igual que con los corrimientos calculados para las señales A1, se puede ver claramente un aumento del corrimiento aproximadamente lineal por ventana calculada. Esto sugiere que el algoritmo fue capaz de detectar partes de los corrimientos generados por el movimiento. Una tendencia similar se puede ver en las señales B2, 3, 4 y 5.

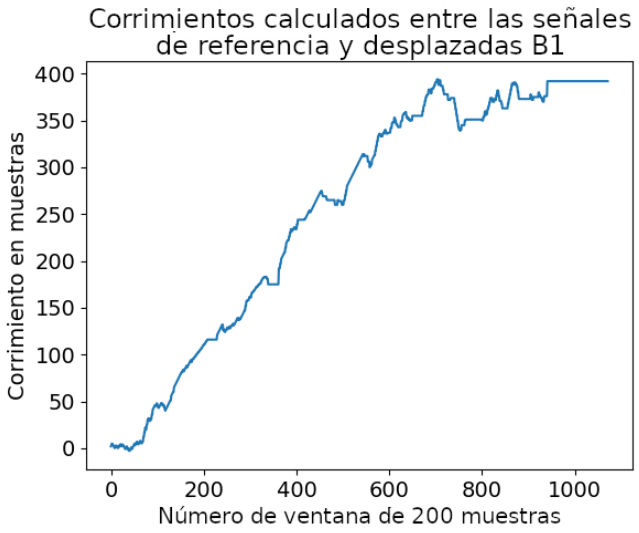

Figura 27. Corrimientos calculados entre las señales de referencia y desplazadas B1.

Al igual que con las señales del grupo A, es posible calcular la velocidad promedio de desplazamiento a partir de la pendiente de los corrimientos. En las señales B1 y B2 se calcula una velocidad promedio de 0.96 m/s y 1.19 m/s respectivamente. Para las señales B3 y B4 los valores de velocidad promedio calculados son 0.55 m/s y 0.76 m/s.

La Figura 28 muestra los corrimientos calculados para el par de señales B3, que tiene dirección negativa. Esto significa que la fuente se encuentra inicialmente cerca del micrófono fijo y se va alejando, lo que resulta, a medida que progresa el audio, que la relación señal a ruido del audio registrado por el micrófono fijo se reduzca. Esta reducción dificulta la correcta detección de los corrimientos. Es por esto que a partir de la ventana número 1000 aproximadamente, los valores de corrimientos se alejan de la línea de tendencia. El mismo patrón se repite para las señales B1, 2, 3 y 4.

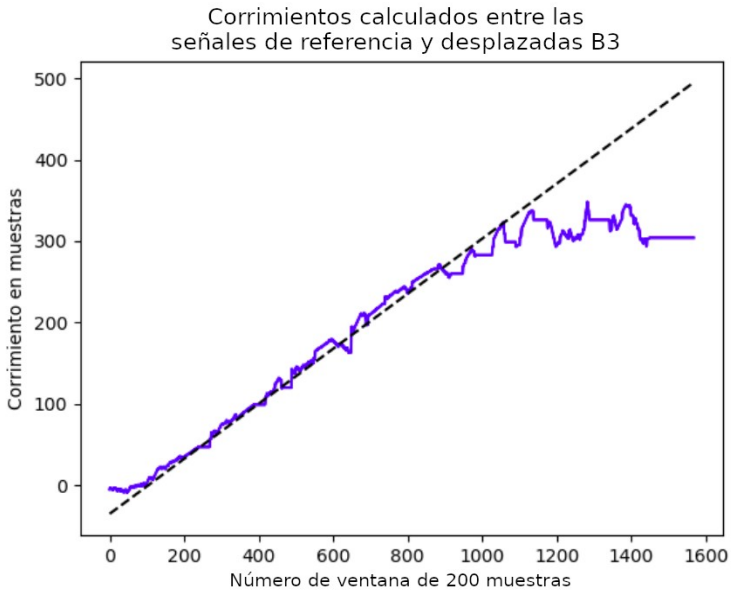

Figura 28. Corrimientos calculados entre las señales de referencia y desplazadas B3 (Azul) y su línea de tendencia (Negra).

La Tabla 4 muestra los valores de correlación entre las diferentes señales B. En primer lugar, la correlación entre la señal de referencia y la desplazada, previo a la alineación. En segundo lugar, la correlación para las señales alineadas con el algoritmo propuesto; por último, la correlación para las señales alineadas usando el software Auto-align Post (AAP).

Tabla 4. Valores de correlación calculados a partir de las señales B de referencia, desplazadas, corregidas y alineadas con el Auto-align Post.

Señal	N°	Descripción	Desalineado	Alineado	Auto-align Post
B	1	Sentido negativo. Velocidad Alta.	0.197	0.490	0.476
	2	Sentido positivo. Velocidad Alta.	-0.116	0.497	0.449
	3	Sentido negativo. Velocidad Baja	0.326	0.622	0.625
	4	Sentido positivo. Velocidad Alta	0.011	0.345	0.340
	5	Micrófono	0.510	0.766	0.757
	6	Rotación derecha	-0.043	0.888	0.868
	7	Rotación izquierda	0.341	0.868	0.853

Dado que estos valores se obtienen correlacionando señales diferentes (mismo evento registrado por dos micrófonos diferentes), no se puede suponer que el valor de

correlación en el caso de una alineación perfecta vaya a ser 1. Por lo tanto, se tomó como referencia el valor de correlación entre la señal de referencia y la señal desplazada alineada con el software AAP. Como se puede ver en los 7 casos, los valores de correlación obtenidos con el algoritmo propuesto son similares o, en algunos casos, ligeramente superiores a los obtenidos con el AAP. Esto sugiere que el algoritmo propuesto fue capaz de identificar y posteriormente alinear las pistas de forma similar al AAP.

En la tabla 5 se muestran los valores de los umbrales optimizados para los pares de señales B.

Tabla 5. Valores de umbrales optimizados para los pares de señales B.

Señal	N°	Descripción	Umbral de amplitud	Umbral de correlación
B	1	Sentido negativo. Velocidad Alta	0.07	0.12
	2	Sentido positivo. Velocidad Alta	0.08	0.05
	3	Sentido negativo. Velocidad Baja	0.10	0.10
	4	Sentido positivo. Velocidad Baja	0.09	0.11
	5	Micrófono	0.07	0.05
	6	Rotación derecha	0.06	0.08
	7	Rotación izquierda	0.05	0.03

La Figura 29 muestra el gráfico de las señales de referencia y corregidas B1 por el algoritmo propuesto y por el AAP. Se puede ver que la corrección del AAP (negra) y del algoritmo propuesto (amarilla) son similares. La mayor diferencia entre la señal corregida usando el algoritmo propuesto y usando el AAP ocurre cuando una de las señales a alinear tiene baja relación señal a ruido. En la Figura 30 se muestra una comparación entre las señales B3 para un caso de baja relación señal a ruido. Como se puede observar, deja de haber coincidencia entre la señal alineada con el algoritmo propuesto y con el AAP.

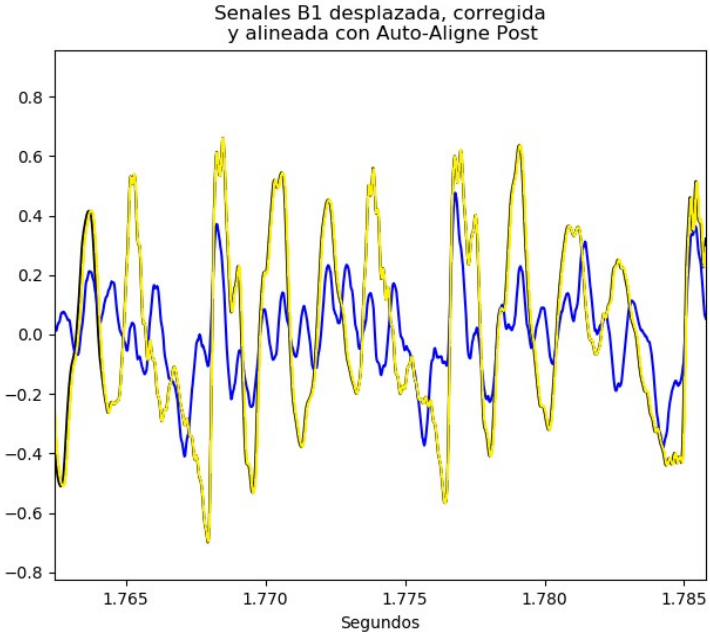

Figura 29. Señales B1 de referencia (Azul), corregida (Amarilla) y Alineada con Auto-align Post (Negra).

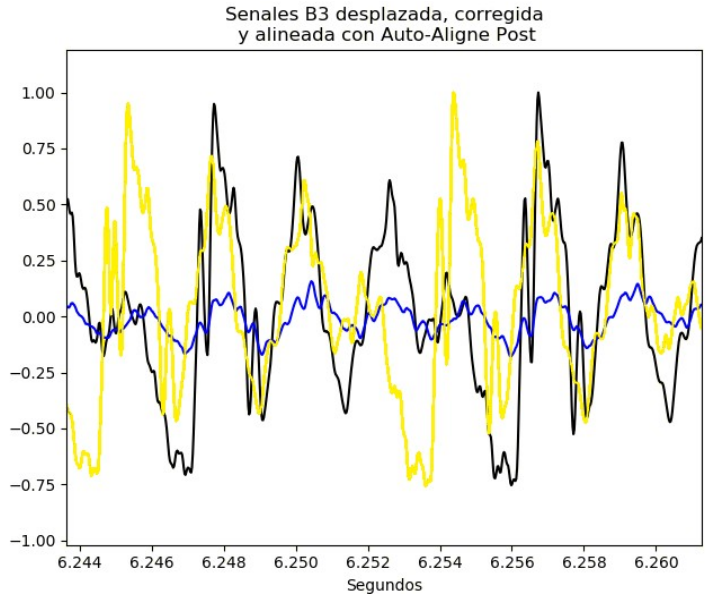

Figura 30. Señales B3 de referencia (Azul), corregida (Amarilla) y Alineada con Auto-align Post (Negra). Situación de baja relación señal a ruido.

Las cancelaciones y filtros peine se ven menos definidas en los espectrogramas de las señales B que en las A. De todas formas, es posible distinguir algunas curvas que desaparecen luego de procesar la señal. La Figura 31 muestra un espectrograma de la mezcla entre las señales B1 de referencia y desplazada y entre las señales de referencia y corregida.

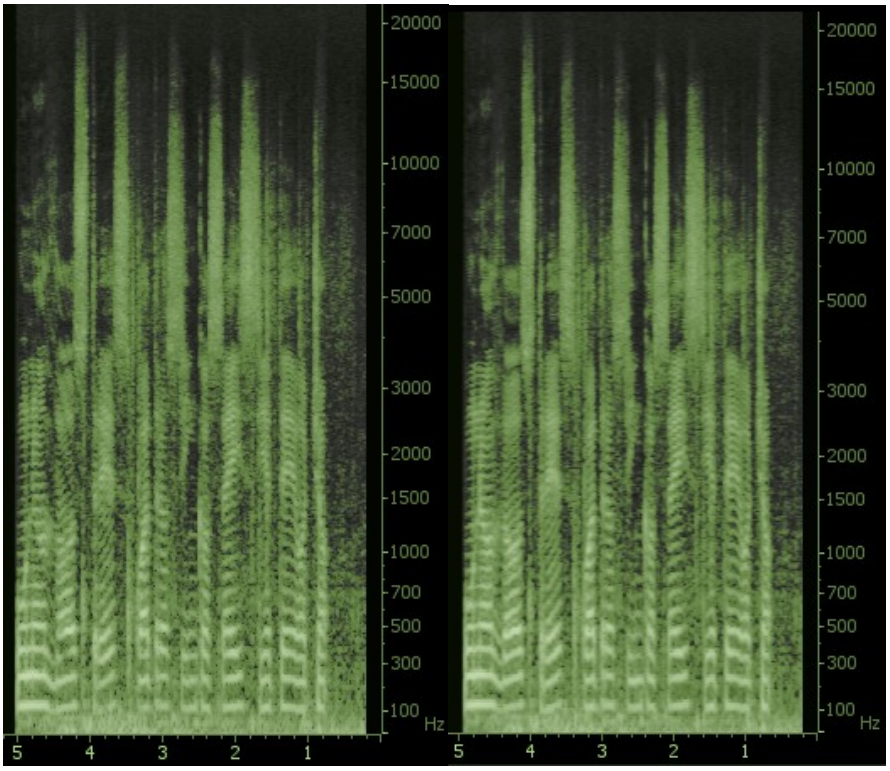

Figura 31. Espectrogramas de la mezcla entre la señal de referencia y desplazada (Izquierda) y corregida (Derecha) B1.

A partir de lo anterior, se puede concluir que el algoritmo propuesto fue capaz de identificar y corregir en gran medida los corrimientos entre las pistas del grupo B.

6.3. SEÑALES C

El grupo de señales C se diferencia de los dos grupos anteriores en que se desconoce por completo el tipo de movimiento que genera el corrimiento entre las pistas. Lo único que es posible predecir de los corrimientos es que los valores son mucho menores que para los dos grupos anteriores. Esto se puede observar en la Figura 32 correspondiente a los corrimientos calculados para el par de señales C2, donde el máximo corrimiento calculado para una ventana es de 26 muestras.

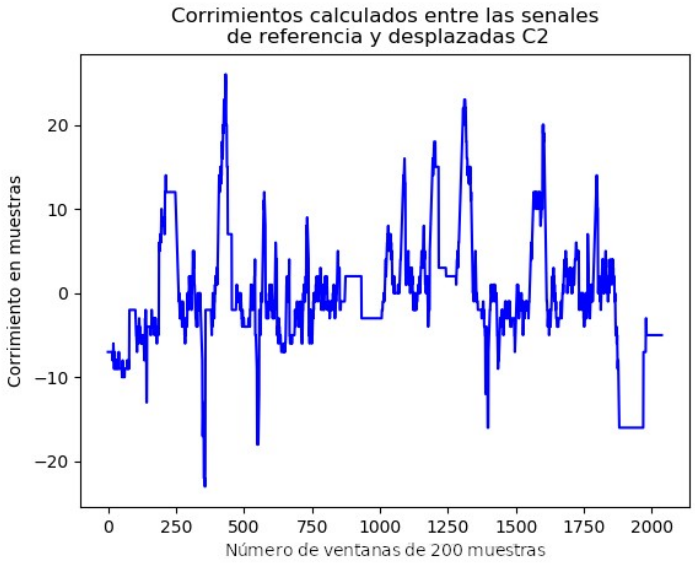

Figura 32. Corrimientos calculados para el par de señales C2

Esta tendencia no se mantiene en los valores de corrimientos entre el par de señales C1. En este caso, el máximo valor de corrimiento por ventana supera las 350 muestras. En la Tabla 6 se puede ver como para el par C1, el valor de correlación obtenido con el algoritmo propuesto es incluso menor que con las señales desalineadas. En el caso de las señales corregidas utilizando el Auto-align Post el valor de correlación total no presenta una variación significativa con la señal desalineada. Esto no significa que este programa no sea efectivo para eliminar los corrimientos. Estos pueden ser muy pequeños como para que se perciba una diferencia significativa en el valor de correlación.

Tabla 6. Comparación de valores de correlación para los pares de señales C

Señal	N°	Descripción	Desalineado	Alineado	Auto-align Post
C	1	Toma 1	0.545	0.395	0.540
	2	Toma 2	0.770	0.793	0.783

La Tabla 7 muestra los valores de umbrales optimizados para las señales C.

Tabla 7. Valores optimizados de umbrales para los pares de señales C.

Señal	N°	Descripción	Umbral de amplitud	Umbral de correlación
C	1	Toma 1	0.03	0.14
	2	Toma 2	0.02	0.05

Esto puede deberse a que en las señales C1 hay 1.5 segundos de silencio en la mitad de la grabación. Durante esos 1.5 segundos, el umbral de amplitud queda por encima de la amplitud de las ventanas, anulando el cálculo de corrimientos por ese tiempo. La Figura 33 muestra la señal C1 donde se ve el intervalo silencioso.

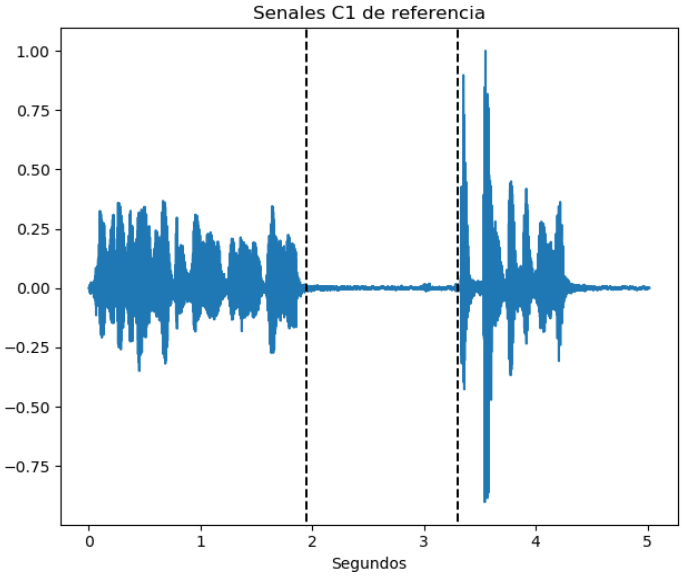

Figura 33. Forma de onda de la señal de referencia C1.

Para verificar esto, se dividió la señal en dos partes: una primera parte que incluye las muestras desde el inicio hasta el segundo 2 y una segunda parte que incluye las muestras desde el segundo 3 hasta el final. La Tabla 8 muestra los valores de correlación calculados a partir de esa división. Como se puede ver, la correlación aumenta significativamente cuando se elimina el segmento silencioso. A su vez, los valores de corrimiento se reducen a valores esperables para el tipo de movimiento del grupo de señales C.

Tabla 8. Valores de correlación de ambas partes de la señal C1.

Señal	Segmento	Desalineado	Alineado	AAP
C1	1	0.614	0.646	0.615
	2	0.502	0.544	0.495

La Tabla 9 muestra los valores de umbrales optimizados para ambos segmentos de la señal C1.

Tabla 9. Valores optimizados de umbrales para los pares de señales C1.

Señal	Segmento	Umbral de amplitud	Umbral de correlación
C1	1	0.11	0.14
	2	0.02	0.14

7. CONCLUSIONES

A partir del análisis de los resultados obtenidos, se puede concluir que el algoritmo propuesto fue capaz de detectar y corregir en gran medida los corrimientos generados por los diferentes movimientos evaluados. Excepto en el grupo C1, los valores de correlación entre las señales fueron mayores luego de aplicar el algoritmo. Además, los valores obtenidos fueron similares o en varios casos superiores a los valores obtenidos usando el AAP. A partir del análisis del espectrograma de la mezcla entre señales, se puede observar que los filtros peine generados por la suma de señales desalineadas, desaparecen en gran medida luego de realizar el procesamiento.

Los resultados obtenidos a partir del procesamiento del grupo de señales C1 sugiere que silencios muy prolongados pueden generar fallas en el algoritmo detector de corrimientos. Al eliminar el silencio, los valores de correlación fueron superiores a los obtenidos con el AAP.

La velocidad de procesamiento fue más lenta que las soluciones comerciales (AAP); en promedio para el algoritmo propuesto se requiere un orden magnitud más.

8. TRABAJOS FUTUROS

A partir del análisis de los resultados obtenidos, se desprenden varios posibles trabajos futuros y mejoras al algoritmo.

- Incorporar una etapa de detección de silencios, que como se vio en los resultados, puede mejorar mucho el desempeño del algoritmo.
- Evaluar el desempeño del algoritmo con señales de audio con otros corrimientos y velocidades de movimiento de fuente, así como también usando diferentes frecuencias de muestreo y profundidades de bit.
- Realizar *tests* de pruebas subjetivas.
- Incorporar una interfaz gráfica que permita un uso más amigable en producciones audiovisuales.
- Posibilitar su utilización como *Plug-in* de las DAW y software de edición para producciones audiovisuales.
- Optimizar el código para reducir tiempos de procesamiento.
- Incorporar la posibilidad de tener más de dos señales de entrada.

9. REFERENCIAS

[1] Purcell, J., Dialogue editing for motion pictures, 2da edición, Routledge, Londres (2015)

[2] Goussios C. A., Sevastiadis C. V. and Kalliris G. M., Outdoor and Indoor Recording for Motion Picture. A comparative approach on microphone techniques, presented at the 122nd AES conv., (2007)

[3] Rumsey, F. “Audio for cinema. Dialog recording and future trends”. J. Audio Eng. Soc., Vol. 66, No. 3, (2018)

[4] Sound Radix, Auto-Align Post, Extraído el 27 de octubre del 2020, shorturl.at/kqsOY

[5] Izotope, Mixing Audio for Video, Part 2: Audio Post-Production Workflows, Extraído el 27 de octubre del 2020, shorturl.at/lovD8 shorturl.at/lovD8

[6] Backtracks, Boom Microphone, Extraído el 27 de octubre del 2020, https://backtracks.fm/resources/podcast-dictionary/boom+microphone

[7] Shure, Choosing a Shotgun Microphone: The Long and Short of It, Extraído el 27 de octubre del 2020, shorturl.at/hrAO0

[8] Sennheiser, MKH 60 P 48, Extraído el 27 de octubre del 2020, shorturl.at/swzBV .

[9] Grant Tony, Audio for single camera operation, Focal Press, Oxford, (2003).

[10] Sanken, COS11, Extraído el 27 de octubre del 2020, shorturl.at/aRS47

[11] Waves, InPhase, Extraído el 27 de octubre del 2020, shorturl.at/jpuP5

[12] Izotope, Phase, Extraído el 27 de octubre del 2020, shorturl.at/gopGI

[13] Adobe, Adobe Audition. Una estación de trabajo de audio profesional, Extraído el 27 de octubre del 2020, shorturl.at/orLNY

[14] Red Giant, Pluraleyes 4, Extraído el 27 de octubre del 2020, shorturl.at/gFW68.

[15] European Broadcasting Union. “The relative timing of the sound and vision components of a television signal”, EBU Recommendation R37-2007, (2007)

[16] David M. Howard, Jaime A. S. Angus. “Acoustics and Psychoacoustics”

[17] Long, M., Architectural Acoustics, 2da edición, Academic Press, Burlington, MA 01803, USA (2014)

[18] Chakraborty, S., Advantages of Blackman Window over Hamming Window Method for designing FIR Filter, (2013)

[19] Titze, I.R., Principles of Voice Production, Prentice Hall (currently published by NCVS.org) (pp. 188), ISBN 978-0-13-717893-3 (1994)

[20] Stevens, K. N., Acoustic Phonetics. Cambridge, MA: The MIT Press, (1998)

[21] John G. Proakis, Dimitris G. Manolakis, Digital Signal Processing, Principles, Algorithms, and Applications, Third edition, Prentice-Hall International, INC., New Jersey (1996) .

[22] Scipy, scipy.signal.firwin, Extraído el 27 de octubre del 2020, shorturl.at/hjxzH.

[23] Jeffs, R., Evolution of the DJ Mixer Crossfader, RaneNote 146, (1999) .

[24] Scipy, numpy.correlate, Extraído el 27 de octubre del 2020, shorturl.at/uzD49

[25] Freesound, Voice Request #19b - This is a story of Liam Timmons.wav, Extraído el 27 de octubre del 2020, https://freesound.org/s/384272/

[26] VietmusicHouse, Extraído el 27 de octubre del 2020, https://vietmusichouse.com/

[27] Ñandú Sonido, Extraído el 27 de octubre del 2020, http://nandusonido.com/