



# UAI

**Universidad Abierta  
Interamericana**

## **Sincronización Multiusuario Para Transmisión en Directo de Obras Musicalizadas**

Tutor de trabajo final: Fernando Asteasuain

Profesora de trabajo final: Marcela Samela

Nahuel Hernán Patera

Trabajo Final de Carrera presentado para obtener el título de  
Lic. en Gestión de Tecnología informática

Diciembre, 2022

---

---

## Resumen

En este Trabajo de Investigación se analizaron los distintos tipos de sincronización de archivos multimedia, y la manera en que afectaban a la sincronización multiusuario para transmisión en directo de obras musicalizadas. Este tipo de transmisión implicaba que, al menos dos personas, pudiesen realizar transmisiones en el mismo momento, pero desde distintos lugares físicos, y con distintos entornos de red.

Era necesario, además, que todas las transmisiones pudiesen ser oídas en perfecta sincronía por terceros, que cumplieran únicamente el rol de receptores. Éstos no debieron haber notado ninguna diferencia temporal entre las transmisiones, pese a que cada entorno de red emisor puede tener distintos tiempos de carga. La idea fue que pudieran vivenciar la experiencia en directo tal como la hubiesen percibido en vivo, con posibles errores humanos, pero sin errores tecnológicos.

En base a las conclusiones obtenidas de nuestra investigación, se propuso resolver el problema con una transmisión levemente diferida, en base a un cálculo a partir del tiempo mínimo de duración de un segmento multimedia. Se contempló también que la transmisión de cada terminal pueda modificarse dinámicamente, en cuanto a calidad y cantidad de resoluciones. En última instancia, si una terminal presentaba demoras muy grandes, solo transmitiría audio. Todo esto minimizó la posibilidad de retrasos mayores a los aceptados por el servidor.

La propuesta contempló la transmisión entre los emisores y el servidor, la sincronización entre ellos y la diversificación de resoluciones para satisfacer las demandas de diferentes consumidores. La arquitectura propuesta en este trabajo resolvió un problema común a varias disciplinas y resultó aplicable a cualquier transmisión audiovisual similar. Precisamente, a cualquiera en la que a sus participantes emisores les haya resultado suficiente, para realizarla, la utilización de una pista guía base.

***Palabras clave:*** audiovisual, multimedia, multiusuario, sincronización, transmisión

---

### **Dedicatoria**

Dedico este trabajo final a cada una de las personas que confiaron en que, algún día, sin importar el tiempo que me llevase, lograría cumplir el objetivo que me había planteado y que con él concluye. A toda mi familia, a mi novia, y a compañeros de estudio y de trabajo, tanto actuales como pasados. Especialmente por los momentos en que me incentivaron, me animaron y me acompañaron. Sobre todo, a mis familiares más cercanos y a mi novia, quienes vivieron más de cerca este proceso, con los tiempos, formas y momentos que dediqué, y me apoyaron en cada decisión.

---

### **Reconocimientos**

Agradezco enormemente a Dios por cada una de las oportunidades y situaciones que ha puesto en mi vida, y las que ha quitado, para que pueda llegar a este momento, reconociendo que nada podría haber logrado si no fuese su voluntad. A todas las personas que contribuyeron a mi educación a lo largo de mi vida, mis padres, abuelos, tíos, padrinos, y cada docente que me ha formado. Desde la educación primaria y secundaria, en el Instituto Educativo Doctor Alexander Fleming, polimodal en el Instituto Manuel Belgrano, terciaria en el Centro Educativo Loreto y universitaria en la Universidad Abierta Interamericana, han contribuido a que pueda lograr mis objetivos académicos.

Agradezco particularmente por el desarrollo de este trabajo a mi compañero de José Marcelo D'Agostino, quien, algunas semanas después de haber escuchado en clases la propuesta de mi trabajo final, me comentó haber escuchado un programa de radio que podía interesarme. Del mismo pude escuchar una grabación publicada en el sitio oficial de dicha radio y se convirtió en el mencionado como caso Disney, dentro del capítulo 2. Por otro lado, pese a que existe bibliografía que indica que es de mal gusto hacerlo, no puedo dejar de agradecer a mi tutor, Fernando Asteasuain, por el acompañamiento, guía y predisposición a ayudarme durante este proceso.

---

## Índice General

### Contenido

Resumen.....	1
<i>Palabras clave</i> .....	1
Dedicatoria.....	2
Reconocimientos.....	3
Índice General.....	4
Índice de Gráficos.....	7
Índice de Tablas .....	8
Capítulo 1 – Introducción y Naturaleza del Problema.....	9
Introducción .....	9
Identificación del Problema .....	9
Planteamiento del Problema .....	11
Justificación .....	11
Objetivo del TF.....	12
Objetivo General.....	12
Objetivos Específicos.....	12
Hipótesis .....	12
Variables .....	13
Nuestra Propuesta .....	13
Marco Teórico.....	14
Tipos de Sincronización Multimedia .....	14
Marcas de Reloj / Sincronización de Relojes .....	23
Calidad de Servicio y Calidad de Experiencia.....	24
Umbral de Percepción Humana .....	31

---

Umbral de Error Humano .....	35
Margen Tecnológico .....	42
Coros y Orquestas .....	45
Contribuciones Principales .....	50
Estructura General de la Tesis .....	50
Capítulo 1 – Introducción y Naturaleza del Problema.....	50
Capítulo 2 – Trabajos Relacionados .....	50
Capítulo 3 – Desarrollo.....	51
Capítulo 4 – Discusión.....	51
Capítulo 2 – Trabajos Relacionados .....	52
Introducción .....	52
Fuentes Relevantes.....	52
Casos Similares.....	54
Capítulo 3 - Desarrollo.....	56
Introducción .....	56
Puntos de Vista .....	56
El Buffer.....	58
Buffer Dinámico .....	59
¿Dónde Utilizamos el Buffer? .....	61
¿Cuánto Tiempo Almacenar y Qué Impacto Genera? .....	62
Transcodificación.....	67
Sincronización y Marcas Temporales.....	69
Guía Base .....	70
Marcas Temporales.....	70
Sincronización.....	70

---

Pruebas de Entorno .....	72
Resumen.....	73
Capítulo 4 - Discusión .....	74
Introducción .....	74
Flujo Propuesto .....	74
Recomendaciones Extra.....	75
Conclusiones.....	77
Líneas Futuras de Investigación.....	79
Acrónimos.....	80
Referencias.....	83

---

## Índice de Gráficos

Figura 1. Jerarquía multimedia ascendente y descendente. ....	16
Figura 2. Variabilidad de retardo. ....	26
Figura 3. Umbrales de detectabilidad y aceptabilidad humana. ....	34
Figura 4. Desfasaje en golpes ejecutados por un acompañante musical de danza. ....	37
Figura 5. Cuatro capas de sincronización presentes en Huang & Nahrstedt. ....	52
Figura 6. Esquema de solicitudes propuesto. ....	65

---

## Índice de Tablas

Tiempos de ejecución de TIC expresados en milisegundos. ....	38
Golpes certeros, adelantados o atrasados, dentro y fuera de la ventana de tolerancia.....	39
Rango de milisegundos considerado adelantado, cero, atrasado o fuera de rango. ....	39
Desvíos mínimos, máximos, medios y desvíos estándares en la ejecución musical. ....	40
Demora aproximada agregada por distintos micrófonos. ....	43
Demora aproximada agregada por distintos altavoces, en función a la distancia.....	43

---

## Capítulo 1 – Introducción y Naturaleza del Problema

### Introducción

Este primer capítulo nos introduce al problema planteado en este Trabajo de Investigación, explicando el mismo y su necesidad de solución, debida a su aplicación en casos reales que lo han motivado. Se pretende que el lector tenga un panorama general sobre el problema en cuestión y los objetivos que han sido planteados.

### Identificación del Problema

Se han podido experimentar, en los últimos años, diversas innovaciones para la interconectividad de individuos aislados, potenciadas y motivadas en muchos casos por la pandemia de COVID-19. La mencionada Pandemia motivó el aislamiento de personas en una gran cantidad de países a nivel mundial, a partir de principios del año 2020. Sin embargo, aún existen algunas actividades específicas que resultaron irremplazables o, al menos, impracticables por medios electrónicos.

Una actividad que sufre aún esta necesidad de innovación, para poder tener una correcta performance, es la de las prácticas corales y orquestales, tanto para ensayos como para presentaciones. La correcta ejecución integrada de estas prácticas requiere el oír en sincronía temporal a cada participante de manera conjunta. Esto se ha tornado imposible de realizar por medio de las conocidas aplicaciones de videollamadas, debido a las limitaciones de red y tecnología existentes. No solo se trata de una limitación por el avance actual de las redes disponibles en el mercado, sino por las diferentes posibilidades de acceso a ellas. Tanto por factores económicos como demográficos, un individuo puede tener acceso a una red de características más limitadas que otro.

Quienes intentan realizar prácticas corales u orquestales por medio de videollamadas, experimentan desfasajes de tiempo que se retroalimentan los unos a los otros, produciendo que el problema crezca mientras transcurre la actividad. Cada músico suele escuchar que el sonido producido por el resto es posterior al que él mismo está produciendo. Esto produce una tendencia a retardar sus notas para acoplarse al resto, sin ser consciente de que empeora la situación. Posiblemente, todos los integrantes hayan comenzado la obra más sincronizados de lo que creen, pero las velocidades de carga y descarga de cada uno producen un retraso temporal. Esta demora,

---

incluso, suele ser distinta para cada participante, ya que el entorno de red de cada uno es diferente.

Se tiene también en consideración la figura del director como orquestador del tiempo de ejecución de cada nota de la obra. El director de un coro o una orquesta, ejecuta una serie de señas con sus brazos y manos, para que cada músico o corista sepa cuándo y cómo intervenir con sus aportes musicales. Es posible que todos los músicos se guíen únicamente por estas señas y no por el resto de los sonidos que, si bien podrían ser silenciados, no resulta suficiente solución para el problema. Como mencionamos, los tiempos de carga y descarga de cada terminal son diferentes. Esto produce que ni el director, ni ningún oyente, puedan oír un sonido uniforme o coordinado.

Se plantea, entonces, la necesidad de una herramienta, o tecnología, que permita la sincronización de las obras musicales, resolviendo, o bien evitando, los problemas mencionados. Actualmente, la manera de crear una obra musical conjunta, es realizarla de manera asincrónica, e involucra un trabajo manual de unificación. Los coros y orquestas, suelen dividir las tareas y realizar grabaciones individuales de los aportes de cada integrante. Luego, todos los videos son enviados a una persona, quien estará a cargo del trabajo de edición. Esta persona unifica los videos teniendo en cuenta tanto la imagen como el sonido, buscando, mínimamente, que el primer sonido de cada video se escuche en el momento exacto que se espera escucharlo.

Solo una vez que el trabajo de edición esta realizado, los integrantes de grupo musical tienen la posibilidad de verlo y oírlo, distribuirlo o transmitirlo por algún medio. El abanico de posibilidades audiovisuales existentes no resulta suficiente para los coros y orquestas en la actualidad. El propio presidente de la Asociación Argentina para la Música Coral, ha comentado a Infobae, un conocido diario argentino, lo siguiente:

No hay forma de que nosotros podamos trabajar si no estamos juntos. Eso implica una ruptura esencial con profundo dolor y con profunda inestabilidad en toda nuestra actividad. En este momento ninguno de los cuerpos estables del Teatro Argentino está trabajando. Esto significa una pérdida enorme. (Sáliche & Rodríguez, 2021)

---

## **Planteamiento del Problema**

En el marco de las restricciones que impiden a coros y orquestas ejercer su actividad de manera presencial, se detectó la imposibilidad continuar con la misma. Si bien el problema de fondo es reafirmado y comprobado por dicha restricción, no es consecuente a ella. Es decir, no radica en el no poder realizar las actividades presenciales, sino, en el no tener una herramienta que permita hacerlo de forma remota y distribuida. Tanto para casos forzados por situaciones ajenas a los conjuntos instrumentales, como por posibles deseos de realizarlo de esa forma, nos planteamos la siguiente pregunta. ¿De qué manera podría crearse un entorno en línea que permita, a cualquier conjunto de personas físicamente aisladas, la grabación o transmisión de contenidos audiovisuales sincronizados, sin diferirlos considerablemente en el tiempo?

## **Justificación**

Esta investigación será de utilidad para resolver la necesidad de sincronización de cualquier actividad que comparta ciertas características con la grabación o transmisión de obras musicales de coros y orquestas. Específicamente, cualquier actividad que requiera sincronizar transmisiones de video de varios participantes, sin que sean diferidas en el tiempo de manera considerable. Cualquier actividad en la que cada participante pueda realizar su transmisión apoyándose únicamente en un audio, video, o melodía de fondo, resultará beneficiada por el modelo que propondremos.

Existen algunos trabajos de investigación, que han tratado el tema de la sincronización multimedia, no obstante, se enfocan en entornos de redes de área local, como, por ejemplo, Yilmaz, Tekalp, & Unluturk (2015). En estas redes, la latencia, velocidad o ancho de banda, resultan regulares, medibles y gestionables en mayor medida que en redes más amplias. El sentido de aplicación a gran escala de la solución que pueda proponerse radica, fundamentalmente, en su aplicación a redes que crucen la frontera interna, siendo incluso redes de alcance mundial (WAN). Este factor puede ser determinante para la viabilidad de proyectos.

Una plataforma basada en nuestra propuesta podría ser extendida al entorno comercial dirigido a espectadores, por medio de transmisiones en línea en tiempo real. Esto representaría una oferta mucho más real que las transmisiones diferidas de videos editados, donde cada parte pudo haberse grabado en días diferentes, varias veces o corregidos sus defectos. Se evita la necesidad de edición manual y se otorga el valor agregado del realismo, dejando lugar a posibles errores humanos y resaltando así el mérito de los intérpretes.

---

## **Objetivo del TF**

### ***Objetivo General***

Diseñar una arquitectura para la sincronización automática de contenidos audiovisuales transmitidos por distintas personas, en distintos lugares físicos y con entornos de red diferentes, con la cual se pueda transmitir una obra conjunta sin desfasajes temporales producidos por la tecnología.

### ***Objetivos Específicos***

Comprender cada uno de los tipos de sincronización de contenido audiovisual, y cuáles son importantes para la sincronización de obras musicalizadas.

Identificar los parámetros de calidad de experiencia necesarios para mantener la percepción del trabajo unificado de la misma forma que en un ambiente presencial.

Explicar de qué manera las marcas temporales pueden ayudar a la tarea de sincronización.

Proponer una arquitectura de recepción y sincronización que se adapte a los distintos entornos de red de los transmisores.

## **Hipótesis**

El planteo de una herramienta que pueda utilizarse por personas físicamente aisladas conlleva las distintas realidades que existan entre ellas. Una propuesta que involucre tecnologías de última generación no resultaría inclusiva, por lo tanto, solo sería útil para una muestra poblacional reducida. Es necesario tener en cuenta, entonces, no solo las tecnologías e infraestructuras existentes, sino, las posibilidades de acceso a ellas que posean los distintos grupos sociales y sectores demográficos que así lo requieran. Igualmente, está claro que, al hablar de una herramienta que involucra hardware e infraestructura de redes, se puede tender a agrandar el público objetivo, pero será necesario el acceso a dichos recursos, al menos en forma básica, para poder utilizarla. Teniendo en cuenta, incluso, los factores desfavorables del marco mencionado, planteamos la siguiente hipótesis. La transmisión en línea de contenidos audiovisuales sincronizados, emitidos por distintas personas, desde diversos lugares físicos, puede realizarse sin diferirlos considerablemente en el tiempo.

---

## **Variables**

La ejecución de obras musicalizadas debe ser estudiada para comprender las necesidades reales de los grupos que las realizan. Conocer estas necesidades no solo nos permitirá brindar soluciones a ellas, sino, apreciar los parámetros de calidad necesarios, los cuáles también deben ser tenidos en cuenta.

A demás de poner foco en las mencionadas variables, relacionadas al problema a abordar, también debemos hacerlo en las que están más ligadas a la solución del mismo. La sincronización de archivos multimedia será la principal de ellas, ya que una transmisión debe cumplir con parámetros aceptables de sincronía para que pueda brindar una experiencia agradable. En cuanto a dicha transmisión, debe tenerse en cuenta tanto el que sea multiusuario, como las posibilidades de realizarla en tiempo real o en forma diferida.

## **Nuestra Propuesta**

En este trabajo ofrecemos una alternativa donde se centraliza la recepción de transmisiones de todos los proveedores de video en vivo en un mismo servidor. Este mismo se encarga de gestionar las peticiones de los flujos de medios, cuantitativa y cualitativamente, basándose en las posibilidades de transmisión que el entorno de red ofrezca con cada proveedor. Teniendo en cuenta que dicho entorno es inevitablemente cambiante, se gestiona por medio de las solicitudes un buffer que garantiza la disponibilidad de contenido para enviar a los consumidores finales.

La sincronización de los archivos es realizada por medio de marcas temporales que cada proveedor replica a partir de las recibidas en una pista base, sobre la cual se ejecutan las obras o piezas musicales. No se adiciona una tarea de corrección temporal en el servidor, ya que estas marcas brindan la información necesaria para la reproducción final. Lo que sí debe garantizar el servidor es una cantidad de resoluciones tal que ayude a minimizar el flujo de descarga necesario por el consumidor.

Los consumidores gestionan a su vez su propio buffer y las solicitudes al servidor, a partir de las cuales comenzarán la reproducción del contenido. Existe de esta forma un retraso de extremo a extremo ejemplificado en algunos segundos, lo cual consideramos que no afecta al objetivo planteado.

---

## Marco Teórico

### *Tipos de Sincronización Multimedia*

En principio debemos diferenciar dos tipos de media, la continua y la estática, tal como lo indican Montagud, Cesar, Boronat, & Jansen (2018, pág. 4). Este trabajo final se basa por naturaleza en la sincronización multimedia continua, ya que su objeto de análisis se corresponde con la transmisión de video. En cambio, la media estática se corresponde con imágenes individuales. En el resto del trabajo, se mencionará la multimedia o sincronización multimedia sin aclarar que se refiere a la continua, y no a la discreta.

Podemos encontrar varios tipos de sincronización de multimedia. Por un lado, se encuentra la clasificación que tienen que ver con cantidad de remitentes y/o receptores. Por otro, la clasificación que refiere al nivel de la relación de sincronización de datos.

En esta segunda clasificación, para poder comprender los distintos tipos de sincronización de multimedia, es necesario entender la jerarquía de un modelo de datos multimedia. Huang & Nahrstedt (2013) han realizado una descripción del modelo. El mismo está compuesto, en manera descendente, desde lo más general hacia lo más específico, por sesiones, paquetes de medios, modalidades de medios, flujos sensoriales y marcos. Como veremos más adelante, por cada componente de esta jerarquía, existen tipos de sincronización dentro de ellos (intra componente) y entre ellos (inter componente).

**Jerarquía Multimedia.** Presentamos esta jerarquía, en este apartado, de manera descendente, de más general a más específico, a fines de dar un orden en el cual estructuraremos el conocimiento. De esta forma, comenzamos con la sesión, como la parte más general de la jerarquía, y finalizamos con los marcos o tramas, como partes mínimas de la misma.

Una sesión (en inglés, *Session*) es donde, al menos dos sitios, comparten una relación a la cual colaboran, es decir que cada interviniente será parte de la misma sesión. Un solo generador de contenido y un consumidor del él son suficientes para que se genere esta relación, sin necesidad de que ambos produzcan, envíen o compartan contenido. Tampoco es necesario que quien lo comparta sea su generador, existiendo la posibilidad de que el contenido se encuentre alojado y accesible en algún sitio.

Cabe mencionar el caso de dos consumidores del mismo contenido, que no colaboran ni se relacionan entre ellos, sino solo con el proveedor del contenido. En este caso, estamos hablando de dos sesiones diferentes. Cada sesión está compuesta por un consumidor y un

---

proveedor. Podríamos hablar de una sesión única si hubiese una relación o colaboración de los consumidores que afecte al otro.

Un paquete de medios (en inglés, *Bundle*) corresponderá siempre a un único sitio de los colaboradores mencionados. Cada sitio reúne dentro de estos paquetes el conjunto de datos multimediales necesario. Dentro de esta jerarquía no se debe confundir la palabra paquete con aquellos datos enviados dentro de los marcos mencionados más adelante. El paquete de medios se refiere al conjunto de modalidades de medios agrupadas en dicho paquete.

Las modalidades de medios (en inglés, *Media Modality*) corresponden a cada uno de los tipos de medios que formaran nuestros multimedios propiamente dichos. Si el lector asistió a una función de cine en 4D, habrá experimentado las modalidades de medios visual, auditiva y háptica. De esta manera, basado en su propia experiencia, podrá comprender el concepto de estas modalidades y la importancia de su sincronización. Todas ellas pueden estar presentes en un mismo paquete.

Los medios poseen a su vez uno o varios flujos sensoriales (en inglés, *Sensory Stream*), cada uno producido por un sensor. Estos sensores podrían ser, por ejemplo, cámaras o micrófonos, si hablamos de un video convencional. Continuando con nuestro ejemplo cinematográfico, es posible que en la modalidad de video se transmitan dos flujos de cámaras diferentes para lograr una imagen en 3D. “Cada sensor produce un flujo sensorial” (Huang & Nahrstedt, 2013, pág. 3), por lo cual, este ejemplo se trata de dos flujos sensoriales de la misma modalidad. Dentro de la modalidad háptica podrían también enviarse flujos de varios sensores, por ejemplo, uno relacionado al tacto y otro al olfato. Completando las modalidades, podríamos encontrarnos con más de un flujo de audio que en conjunto formarían un sonido estéreo.

“Un flujo de video (media continua) se compone de una colección de marcos o imágenes subsecuentes que necesitan ser presentadas en el mismo orden y con la misma duración en que fueron capturados” (Montagud, Cesar, Boronat, & Jansen, 2018, pág. 4).

Como último elemento de esta jerarquía se encuentran los marcos o tramas (en inglés, *Frames*), que son los datos capturados por los sensores, fraccionados y enviados a través de la jerarquía mencionada. Un flujo de video podría componerse, por ejemplo, por 30 o 60 marcos por segundo (en inglés, *frames per second* o FPS). Para ello, el flujo especifica la cantidad de tiempo que cada marco debe ser reproducido. Se los llama también marcos de medios (en inglés,

*media frames*), teniendo en cuenta que las imágenes son solo una modalidad de medios. El sonido se transmite también por medio de marcos y con tiempos de reproducción.

En la siguiente figura podemos apreciar la jerarquía mencionada, agrupada de forma descendente desde el elemento más general hacia el más específico, todos ellos acompañados de sus traducciones. Pretendemos no perder de consideración las palabras en inglés, ya que es la forma en la que se encuentran en la mayor cantidad de bibliografías que consultamos.

**Figura 1.**

*Jerarquía multimedia ascendente y descendente.*



**Cantidad de Remitentes.** Cuando tenemos más de un remitente o fuente de archivos que deben ser sincronizados entre sí, damos lugar a la sincronización entre fuentes, o inter-fuente (en inglés, Inter-Source Synchronization). También es llamada Sincronización entre emisores (en inglés, Inter-Sender Synchronization), o multi-fuente.

Montagud, Boronat, Martínez, Belda, & Cesar (2015) la mencionan como un subtipo específico de la sincronización inter-media, aunque cabe destacar que dicho autor no utiliza toda la jerarquía presentada en este trabajo, sino, solo distingue la jerarquía entre intra-media e inter-media. Cuando esto ocurre, se debe prestar especial atención a no trasladar la misma afirmación a la jerarquía completa, ya que dicha fuente podría incluir la sincronización de paquetes y sesión al hablar de sincronización intra-media.

En MediaSync (Montagud, Cesar, Boronat, & Jansen, 2018, pág. 11), se menciona también que esto corresponde a un subtipo específico de la sincronización Inter-media. Adicionalmente, indica que “un ejemplo típico es la sincronización entre transmisiones de video enviadas por diferentes servidores de video (por ejemplo, en sistemas de vigilancia)” (Montagud, Cesar, Boronat, & Jansen, 2018, pág. 11).

---

Este tipo de sincronización es objeto de estudio de nuestro trabajo de investigación, donde partimos de la necesidad de sincronización de diversas fuentes de contenidos audiovisuales. Se volverá a hacer referencia a esto dentro del apartado de Sincronización Intra Sesión, de la página 22.

**Cantidad de Receptores.** Existen diversos casos en los cuales es necesario que varios receptores visualicen los acontecimientos al mismo tiempo. En estos casos, hablamos de sincronización inter-receptor (en inglés, inter-receiver) o multi-receptor. Este tipo de sincronización es mayormente conocido por las abreviaturas en inglés IDES e IDMS, que se explayarán a continuación.

En este caso, y a diferencia de las fuentes, se suele hablar de dispositivos en lugar de receptores, de esta forma, surge el concepto de sincronización inter-dispositivo (en inglés, Inter-Device Synchronization, conocida como IDES). A su vez, se establece una diferencia entre dispositivos o destinos, dependiendo de la ubicación y la distancia entre ellos. Así surge la sincronización de medios entre destinos (en inglés Inter-Destination Media Synchronization, conocida como IDMS).

En este trabajo no se profundizará técnicamente en como lograr esta sincronización ni en las diferencias entre IDES e IDMS, pero mencionamos a continuación, y modo de ejemplo, la distinción entre ellas comentada por Montagud, Cesar, Boronat, & Jansen, (2018):

Los receptores/dispositivos involucrados que reproducen el mismo contenido o uno relacionado pueden estar físicamente cerca (por ejemplo, en la misma red local) o distribuidos geográficamente (por ejemplo, en diferentes edificios, ciudades o países). Ejemplo de la primera situación son los escenarios multipantalla y edificios con múltiples altavoces distribuidos. En tales casos, la ausencia de IDES dará como resultado experiencias multipantalla incoherentes y la percepción de molestos efectos de eco, respectivamente. Cuando los dispositivos involucrados están muy separados, el término sincronización de medios entre destinos (IDMS) se emplea típicamente para referirse a la sincronización entre sus procesos de reproducción. (pág. 12)

---

Recomendamos a quien esté interesado en profundizar sobre la sincronización entre usuarios, tener en consideración la búsqueda de información sobre sincronización de juegos multijugador en línea. Estos tipos de juegos representan un escenario donde una pequeña diferencia de tiempo puede representar la victoria de un jugador y derrota de otro. Sin ahondar en los ejemplos que podrían desencadenar dicha situación, destacamos su importancia y criticidad.

**Sincronización Intra Flujo (en Inglés, Intra-Stream Synchronization).** La sincronización Intra flujo refiere al ordenamiento temporal de los marcos de medios dentro de los flujos sensoriales. Es importante destacar la palabra *dentro*, ya que no se trata de la sincronización entre distintos flujos. En concordancia, Nilsson (2018) menciona en su definición que se trata del ordenamiento de los marcos de un único flujo de medios. No obstante, aunque no se trate en esta instancia de sincronizar distintos flujos entre sí, este tipo de sincronización debe estar presente en cada uno de ellos.

“La sincronización Intra-Flujo prescribe la presentación sincrónica de marcos de medios dentro de cada flujo sensorial en los receptores, de acuerdo con su línea de tiempo original capturada en los sensores multimedia.” (Huang & Nahrstedt, 2013, pág. 3). Esto denota que no organiza la sincronización entre varios flujos, sino solo dentro de ellos, tratando a cada uno como un único flujo independiente. La definición no escapa del juego de palabras a entender, hablando de Intra (dentro, interno) e Inter (entre, externo). De todas formas, merece la atención cada nivel, ya que esto no quedaría tan claro al leer autores que indican que, por ejemplo, inter-media e inter-flujo son lo mismo.

Montagud, Boronat, Martínez, Belda, & Cesar (2015, pág. 356) mencionan la sincronización intra-flujo y a la sincronización intra-media como un único tipo de sincronización. Dicha afirmación no es invalidada por lo anteriormente mencionado, ya que una modalidad podría tener solo un flujo. Como se evidenciará en el próximo apartado, al hablar de sincronización intra-media, se debe pensar en lo que hay dentro de la modalidad de medios a sincronizar. En caso de encontrarnos solo con un flujo de medios, se deberá procurar entonces la sincronización intra-flujo de dicho flujo, ya que no hay más de uno para sincronizar entre ellos.

En casos donde más de un flujo coexistan dentro de la misma modalidad, debe tenerse en cuenta que estos son en realidad dos tipos de sincronización diferentes.

Estas relaciones temporales pueden sufrir alteraciones durante la cadena de distribución. Para mantenerlas y, si es necesario, restablecerlas, estos últimos autores mencionan la necesidad

---

de mecanismos precisos y adaptativos. Dichos mecanismos pueden ser consultados con profundidad en bibliografías como las de Liu & Zarki (2003, 2005, 2006 y 2010). Dentro del capítulo 3 de nuestro trabajo hablaremos de estos mecanismos, en la sección que comienza en la página 58, El Buffer. Adicionalmente, tanto en las bibliografías mencionadas como en He, Cai, & Zhau (2009), pueden consultarse las fórmulas que miden la calidad de la sincronización.

Los errores de sincronización Intra-Flujo pueden causar distorsiones del contenido multimedia, como, por ejemplo, una imagen que viene avanzando en un video y, repentinamente, retrocede y vuelve a avanzar sin sentido. Estos tirones o sacudidas también podrían tener lugar en audios, lo que haría que una canción tenga una fracción de segundos de acople. Básicamente, un fragmento multimedia que debe reproducirse antes que otro, termina reproduciéndose después, o viceversa, y esto afecta directamente a la percepción humana del contenido.

La recepción desordenada de marcos de medios no tiene por qué ser un error de sincronización en el envío de ellos, sino más bien, producto del jitter (explicado en el apartado de Calidad de servicio, de la página 24) y la pérdida de paquetes.

**Sincronización Intra Media (en Inglés, Intra-Media Synchronization).** También conocida como Inter Flujo (en inglés, Inter-Stream), ya que refiere a la sincronización entre distintos flujos, aunque aún de una misma modalidad de medios. Para hacer efectiva esta sincronización, como se mencionó anteriormente, se necesita que existan al menos dos flujos, de una misma modalidad de medios, representando la misma cosa. No es importante cuál sea la modalidad de medios, sino que sea una misma. Por ejemplo, dos cámaras de video filmando una misma escena con distintos ángulos, o varios micrófonos ubicados en los distintos cuerpos de una batería (instrumento de percusión). La Sincronización intra-media se encargará de que los flujos enviados por cada uno de los sensores de la modalidad correspondiente se encuentren en sintonía.

En algunas ocasiones, podemos encontrar autores hablando de sincronización inter-flujo sin referirse exactamente a la sincronización intra media, si no, a la inter-media. Esto es el caso de Liu & Zarki, al indicar que “Un ejemplo típico de sincronización inter-flujo es la “sincronización de labios” en presentaciones de audio/video” (Liu & Zarki, A Synchronization Control Scheme for Real-Time Streaming Multimedia Applications, 2003). Esto refiere a la sincronización entre el movimiento de los labios y el sonido que es emitido por ellos, lo cual corresponde a dos modalidades diferentes, audio y video.

---

El caso mencionado podría haber sido motivado, al igual que en el caso aclarado en la sincronización intra-flujo, por la falta de flujos diferentes de un mismo medio para sincronizar entre sí. Esto deja lugar a pensar en la sincronización inter-flujo como la sincronización entre los flujos únicos de diferentes modalidades de medios. En caso de existir dos o más flujos de un mismo medio y una misma escena, se debe tener en cuenta la diferencia que aquí estamos marcando.

“La sincronización Intra-Media representa la sincronización de flujos sensoriales de múltiples dispositivos de medios, de la misma modalidad de medios, dentro de un paquete de medios” (Huang & Nahrstedt, 2013, pág. 3). La última aclaración de esta afirmación (dentro de un paquete de medios) puede compararse con nuestra aclaración de *una misma escena*. Cada paquete de medios vendrá de una escena y/o sitio diferente. En nuestro objeto de estudio, dos cantantes ubicados en sitios distantes el uno del otro, enviarán paquetes de medios diferentes, por más que utilicen la misma, o mismas, modalidades de medios dentro.

Como menciona Nilsson (2018), que utiliza a Huang & Nahrstedt (2013) como fuente, y en concordancia con ellos, errores de sincronización de esta capa pueden violar la correlación espacial durante una presentación (por ejemplo, falta de coincidencias visuales entre dos vistas de las mismas imágenes).

Su objetivo principal es que las relaciones temporales entre las UM<sup>1</sup> durante la presentación se parezcan lo más posible a las especificadas durante los procesos de captura/muestreo/recuperación, o incluso codificación, en el lado de la fuente, a pesar de la existencia de retrasos y diferencias de retraso a lo largo de la cadena de distribución de extremo a extremo. (Montagud, Cesar, Boronat, & Jansen, 2018, pág. 8)

**Sincronización Intra Paquete (en Inglés, Intra-Bundle Synchronization).** Cuando hablamos de Sincronización Intra-Paquete, nos referimos a la que se da dentro de los paquetes de medios. No se debe confundir un paquete de medios con el fragmento de datos que viajan dentro de un marco, obtenidos directamente en los sensores. El paquete de medios se refiere al grupo de modalidades que serán agrupadas en esta instancia por su relación temporal y espacial. Para

---

<sup>1</sup> UM refiere a unidades de medios, traducción de Media Units, abreviado en los textos como MUs.

---

evitar este tipo de confusiones, podría bien mantenerse su termino en inglés, bundle. Dentro de este trabajo se ha optado, valiendose en esta aclaración, por utilizar paquete como traducción del termino.

También se la puede llamar Sincronización Inter-Media (en inglés, Inter-Media Synchronization), de hecho, hemos encontrado más autores que lo hacen de esta forma, como herramienta para distinguir la intra-media de la inter-media. Es lo mismo hablar de la sincronización dentro de los paquetes o entre las modalidades de medios, ya que, lo que sucede dentro de los paquetes, es la sincronización entre las modalidades, valga la redundancia.

Al hablar de este tipo de sincronización, podemos valernos del ejemplo de la sincronización de labios, muy nombrado en textos en inglés como *Lipsync*, de la que ya hemos hablado. El ejemplo ha sido utilizado y repetido por diversos autores, como Huang & Nahrstedt (2013), Montagud, Boronat, Martínez, Belda, & Cesar (2015), Nilsson (2018) y Montagud, Cesar, Boronat, & Jansen (2018). En el caso de Montagud, Boronat, Martínez, Belda, & Cesar (2015), utilizan el ejemplo mencionando que la sincronización inter-media, también es conocida como sincronización inter-flujo. Aunque en este caso sí se menciona principalmente como sincronización inter-media, a diferencia de como comentamos que lo realizaron Liu & Zarki (2003). Por su parte, Nilsson (2018) también menciona que puede ser llamada sincronización Inter-Flujo.

“Cuando varios componentes de medios correlacionados están involucrados en un sistema multimedia, las dependencias originales entre sus UM también deben ser preservadas durante la presentación. Este es el objetivo de la sincronización Inter-Media” (Montagud, Cesar, Boronat, & Jansen, 2018, pág. 9). En otras palabras, “se ocupa de preservar las dependencias temporales entre las UM de diferentes componentes multimedia” (Montagud, Boronat, Martínez, Belda, & Cesar, 2015).

Como ejemplo de consecuencias de una mala sincronización de este tipo, intentando no ser repetitivos con la sincronización labial, podemos mencionar la visualización de subtítulos previa o posteriormente a haber escuchado el audio correspondiente a ellos. Para corregir esta falla, suele acortarse o estirarse el tiempo de reproducción del subtítulo. En ocasiones, se termina reproduciendo un subtítulo por tan poco tiempo que resulta imposible de leer. Esto provoca que el consumidor experimente una falla muy notoria que afecta directamente a la calidad de experiencia. Como indican Montagud, Boronat, Martínez, Belda, & Cesar, “puede llegar a ser

---

irritante (mala QoE)” (2015). Este tipo de fallas, pueden generar confusión e “incluso información incorrecta al intentar asociar o correlacionar los datos de los sensores involucrados.” (Montagud, Cesar, Boronat, & Jansen, 2018, pág. 10).

**Sincronización Intra Sesión (en Inglés, Intra-Session Synchronization).** “La sincronización inter-media puede implicar también la sincronización de componentes de medios originados en diferentes fuentes, o entregados por diferentes remitentes.” (Montagud, Cesar, Boronat, & Jansen, 2018, pág. 11). Los autores utilizaron este fragmento para mencionar a la sincronización inter-fuente (en inglés, inter-source synchronization). Nosotros llamaremos esto sincronización Intra Sesión. Dentro de una sesión, se deben sincronizar los paquetes de las distintas fuentes entre sí.

Teniendo en cuenta que cada paquete de medios puede corresponder a distintas fuentes, o ser entregado por distintos remitentes, también podríamos llamarla sincronización inter paquete (en inglés, inter-bundle synchronization). Esto último se coincide con la jerarquía de sincronizaciones presentada en este trabajo. Así también, corresponde la sincronización inter-remitente (en inglés, inter-sender), como la llaman Huang & Nahrstedt (2013), quienes indica que “representa la presentación sincronizada de paquetes de medios de múltiples remitentes en el mismo receptor”.

Indican también Huang & Nahrstedt que “un error de sincronización puede llevar a la confusión del usuario receptor cuando está viendo a los emisores realizando una actividad altamente colaborativa” (2013, pág. 4). Este es el caso que ha motivado el presente trabajo. El alto nivel colaborativo de las obras musicales representa una gran importancia de sincronización entre fuentes, por lo tanto, de sincronización intra-sesión.

Ahora bien, como comenta Nilsson, “La sincronización intra-sesión representa tanto la sincronización inter-emisor como la sincronización inter-receptor. Aunque solo es necesario que una sea utilizada para clasificarlo como sincronización intra-sesión” (2018, pág. 4). De acuerdo con esto, no podríamos definir a la sincronización intra-sesión solo mencionando las múltiples fuentes, sino, que deben tenerse en cuenta los múltiples destinos. Fuera de la definición, sí es suficiente la presencia de solo una para poder afirmar que estamos hablando de este tipo de sincronización. Por lo cual, nuestra investigación tiene la necesidad de esta sincronización, en base a sus múltiples emisores, sin importar si deben o no sincronizarse los receptores.

---

Al hablar de la sincronización inter-receptor, Nilsson (2018) indica que un fracaso al sincronizar puede generar injusticias cuando algunos sitios reciben un beneficio temporal para realizar una actividad. Esto es similar a lo mencionado en el apartado de sincronización de múltiples receptores, de la página 17.

### ***Marcas de Reloj / Sincronización de Relojes***

Como explica Nilsson (2018), dos computadoras diferentes, cada una con su propio reloj, pueden tener un sesgo temporal entre ellas, e incluso una variación diferente en el tiempo. Cada reloj tendrá su propio sesgo y, adicionalmente, podrían tener un correr del tiempo distinto entre ellos. Se denomina deriva del reloj cuando estos no avanzan de forma heterogénea. Dado este fenómeno, de no sincronizarse con frecuencia, la diferencia temporal entre distintos relojes crecerá con el correr del tiempo. Al ser posible encontrar dispositivos con relojes desincronizados, Saini & Ooi (2018, pág. 172) explican que utilizar las marcas de tiempo de estos relojes, en las grabaciones, es un enfoque ingenuo para la sincronización multimedia. Cabe destacar, para no quitar de contexto dicha mención, que estos autores refieren a una sincronización del tipo Intra Sesión, con varias fuentes en simultáneo.

Es necesario que los relojes de los dispositivos tengan una actualización constante, para evitar así, la acumulación de variaciones temporales diferentes entre ellos. De esta manera puede persistir una diferencia entre las marcas de reloj de uno y otro dispositivo en un mismo momento dado, pero no se incrementarán indefinidamente. En otras palabras, habrá un límite de diferencia entre ellas, determinado por su diferencia en un momento dado, la frecuencia de actualización y la deriva de estos relojes. Se debe tener en cuenta que la actualización podría también traer consigo ajustes tardíos dependiendo del protocolo. La demora de estos ajustes puede variar y, al no ser exacta su frecuencia, generar también algún desvarío.

Las muestras de audio y los fotogramas de vídeo grabados por el mismo dispositivo en la misma instancia de tiempo, por otro lado, son garantía de tener la misma marca de tiempo de grabación, ya que se basan en el mismo reloj. (Saini & Ooi, 2018, pág. 172)

Los protocolos que suelen utilizarse para la sincronización de relojes son el Protocolo de tiempo de red (Network Time Protocol, o NTP) y el Protocolo de tiempo de precisión (Precision Time Protocol, o PTP). En base a sus pruebas, Nilsson menciona que “ambos funcionan a un nivel aceptable aproximadamente el 95% del tiempo. Esto no es suficientemente bueno, ya que el

---

usuario notará asincronía el 5% del tiempo” (2018, pág. 29). Adicionalmente, se menciona que los errores pueden superar los 100 ms. Este autor ha intentado demostrar con pruebas la viabilidad de utilizar la transmisión vía GPS para la sincronización del reloj. Si bien no resultó viable, al menos con el GPS conectado mediante USB, y ocasionado por los tiempos de actualización de dicho puerto, queda presentada como opción a analizar para algún caso en particular. En hardwares que no sean diseñados especialmente para dicha aplicación, es probable que se carezca de un GPS incorporado.

### ***Calidad de Servicio y Calidad de Experiencia***

Desde un principio, preferimos centrarnos en la calidad de experiencia, ya que es lo que buscamos brindar, una experiencia en línea percibida de igual manera que una presencial, al menos en cuanto a la sincronización sonora. De todas formas, comenzamos describiendo la calidad de servicios, debido a su impacto y relación con la experiencia. Ambas deben ser tenidas en cuenta, enfocándonos en los objetivos planteados.

**Calidad de Servicio.** Cuando una aplicación depende directamente de un entorno de red, como lo es el internet hoy en día, la calidad de este servicio será el punto de partida de cualquier análisis. Lamentablemente, la calidad de servicio del internet no es constante y tampoco resulta fácilmente predecible. La forma en que esta red mundial trabaja se encuentra supeditada al *mejor esfuerzo*. Esto significa que en todo momento intentará brindar el mejor servicio que sea posible, pero esa posibilidad depende, entre otras cosas, de la oferta y demanda disponible de cada momento. La oferta estaría marcada por los caminos posibles y sus anchos de banda, y la demanda, por cuán utilizados son. Su conjunción resulta en la disponibilidad para el envío de datos. En la página 26 se especifican algunos factores influyentes en esta oferta mencionada.

Lo antes descrito produce, no solo ciertos retardos (en inglés, *delay*) inevitables, sino una variabilidad entre ellos, lo que significa que no todos los datos llegaran a destino con el mismo retardo. De esta manera, la recepción de un paquete enviado con anterioridad a otro podría recibirse, incluso, luego de la recepción de este último. Esta fluctuación de la red es mayormente conocida por su término en inglés, *Jitter*. “los retardos y su variabilidad, ya sea para un mismo flujo (*jitter*), entre diferentes flujos o entre diferentes receptores, son los factores más determinantes por los que se van a necesitar soluciones de sincronización multimedia” (Montagud, Boronat, Martínez, Belda, & Cesar, 2015, pág. 3).

---

Cuando hablamos de la calidad de servicio, mayormente conocida como QoS (Quality of Service), por sus siglas en inglés, no hablamos solo de los retardos y su variabilidad. En nuestro objeto de estudio, por ejemplo, debemos hacer foco en la calidad de la sincronización multimedia que, siendo un servicio que prestaremos, es un factor de QoS importante para nosotros. Por lo tanto, podemos decir que tendremos dos requisitos de QoS principales, los requisitos de sincronización y los requisitos de retardo. Como concluyen He, Cai, & Zhau,

La sincronización de bajo nivel de medios de transmisión sigue siendo una tarea desafiante, debido a la red de transporte de "mejor esfuerzo", y de aplicación complicada.

Los comportamientos del canal impredecibles hacen que sea bastante difícil cumplir tanto con los requisitos de sincronización como con los requisitos de retardo simultáneamente.

(2009, pág. 414).

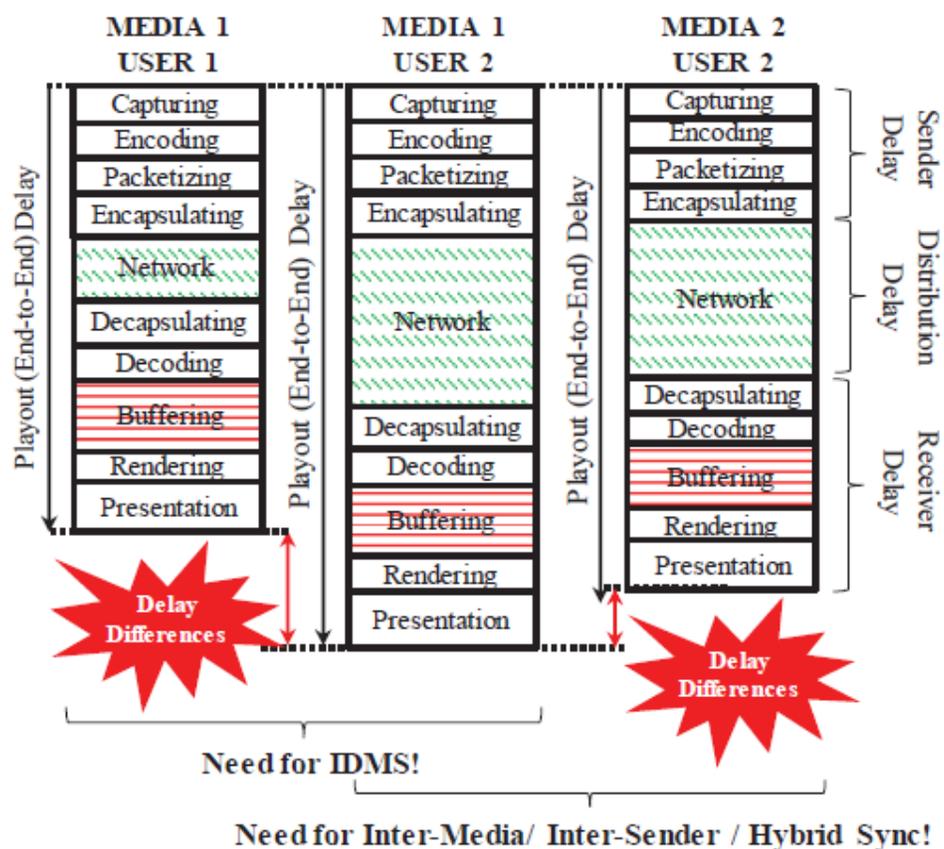
Si bien los mencionados son los aspectos de QoS en los que nos resulta interesante ahondar en esta sección del trabajo, existen otros, como la tasa de pérdidas, propia del canal de distribución, o el ancho de banda del servicio contratado por cada parte. Este ancho de banda no resulta previamente estimable en nuestro escenario, ya que la propuesta debe ser lo más inclusiva que resulte técnicamente posible.

Ahora bien, la red no es el único factor que afecta a los retardos y su variabilidad. Como se puede observar en la Figura 2, podemos dividir los motivadores del retardo en tres partes, el emisor, la red de distribución (internet en nuestro caso de estudio) y el receptor. En la misma, adicionalmente, se pueden observar tanto diferencias de retardo entre la recepción de un mismo paquete por dos usuarios distintos, como en la recepción de dos paquetes distintos por un mismo usuario. Para mayor entendimiento, es preciso detallar algunos factores que afectan dentro de cada una de las tres partes mencionadas.

Dentro del emisor encontramos “procesos como la captura, codificación, cifrado, buffering, paquetización y encapsulamiento, entre otros” (Montagud, Boronat, Martínez, Belda, & Cesar, 2015, pág. 3).

**Figura 2.**

*Variabilidad de retardo.*



*Nota.* Tomado de *Variabilidad de retardos: Necesidad de sincronización multimedia* (pág. 3) por Montagud, Boronat, Martínez, Belda, & Cesar, 2015, Impacto de Parámetros de QoS en Aspectos de QoE: Análisis desde el Punto de Vista de la Sincronización Multimedia.

Dentro de la red de distribución, sumando especificidad a lo mencionado en la página 24, Encontramos:

los procesos de propagación y transmisión de paquetes, tanto a través de los enlaces de comunicaciones como de los dispositivos intermedios y finales involucrados, los procesos de buffering y control en los routers intermedios (p.ej., decisiones de encaminamiento, políticas de QoS...), así como otros procesos avanzados como son la fragmentación y re ensamblado de paquetes, trans-codificación, conversión de formatos,

---

etc. Además, la carga instantánea de la red va a influir significativamente en la magnitud de estos retardos y de su variabilidad, tanto para un mismo flujo multimedia (es decir, el jitter), como entre diferentes flujos multimedia recibidos por el mismo o diferentes dispositivos. (Montagud, Boronat, Martínez, Belda, & Cesar, 2015, pág. 3)

En el receptor, tendremos “procesos de buffering, descifrado, decodificación, de-paquetización o procesado en los diferentes niveles de la pila de protocolos, técnicas de detección y corrección de errores, retardos de entrega (rendering delay), etc.” (Montagud, Boronat, Martínez, Belda, & Cesar, 2015, pág. 3)

Tanto en el lado del emisor como en el del receptor, también habrá una variabilidad de retardos marcada por las capacidades del hardware y software utilizados, teniendo en cuenta también, qué otra cantidad de procesos deben coexistir en cada caso particular. Un factor extra, como comenta la siguiente cita, está relacionado con los relojes, no obstante, en nuestro trabajo se encuentra mayor detalle de estas influencias en el apartado de la página 23, Marcas de Reloj / Sincronización de Relojes.

Otro factor muy importante que puede contribuir a las diferencias de retardos, tanto para un mismo flujo como entre varios flujos, y tanto para un mismo receptor como entre diferentes receptores, son las imperfecciones de los relojes, pues van a acarrear desviaciones (lineales y/o aleatorias) en las tasas de generación, transmisión y reproducción de los contenidos multimedia. (Montagud, Boronat, Martínez, Belda, & Cesar, 2015, pág. 4)

Los retardos ocasionados en el emisor y el receptor se encuentran relacionados estrechamente con la necesidad de sincronización del nivel de producción y presentación, respectivamente. Introducimos aquí el concepto de sesgo como la diferencia cuantificable de sincronización, o nivel de desincronización. La calidad de servicio es un factor medible, y su medición se basa en parámetros. Para estos parámetros, se especifican valores aceptables, por ejemplo, en cantidad de milisegundos de desincronización de labios permitida, tanto para audio antes del video como viceversa.

---

La sincronización del nivel de producción será el primer punto en un circuito de QoS, ya que se sitúa en el lugar y momento donde se produce el contenido. “Por lo general, implica la grabación de datos sincronizados para su posterior reproducción. Los datos almacenados deben ser capturados y registrados sin ningún sesgo, es decir, ‘sincronizados’” (Steinmetz, Human Perception of Jitter and Media Synchronization, 1996, pág. 68). Como mencionamos en la página 27, habrá hardware y software involucrados en la grabación de este contenido. Dos sensores, como un micrófono y una cámara, pueden tener distintos tiempos de retardo en la entrega del contenido que están percibiendo, ello conlleva una desviación del tiempo real. Una computadora puede ser más, o menos, veloz que otra para procesar ese contenido, lo cual habría que dos computadoras que se encuentran grabando el mismo contenido indiquen que fue grabado en momentos diferentes (con diferencias de unos pocos milisegundos). Estas diferencias, entre otras, deben subsanarse desde el mismo generador de contenido, antes de ser enviadas a la red de distribución. “En general, todos los datos sincronizados que se procesarán posteriormente deben sincronizarse de acuerdo con la calidad del nivel de producción, es decir, sin ningún sesgo” (Steinmetz, Human Perception of Jitter and Media Synchronization, 1996, pág. 68).

La sincronización del nivel de presentación será el último punto del circuito de QoS, ya que se sitúa en el lugar y momento donde debe visualizarse el contenido, previamente generado y transportado. Para este nivel, suelen especificarse valores aceptables de sesgo, basados en el análisis de la percepción humana, de la cual hablamos en la página 31. Tal como en el nivel de producción, aquí también podrían presentarse diferencias por el hardware y software utilizados. Por ejemplo, un parlante conectado a una computadora por cable Jack o USB tendrá un tiempo de respuesta diferente al de uno conectado por WIFI o Bluetooth. Existen también monitores con tiempos de respuesta y de actualización diferentes, así también como el tipo de conexión entre ellos y otro dispositivo. “El display lag puede originar una asincronía entre la señal de video y la de audio, afectando, por tanto, a la sincronización inter-media (si no se compensa de algún modo)” (Montagud, Boronat, Martínez, Belda, & Cesar, 2015, pág. 4). Es importante que la información se encuentre previamente sincronizada, o hayan sido añadidos los metadatos necesarios para su posterior reproducción sincronizada en este nivel, ya que en la presentación pueden ser aún más influyentes los sesgos introducidos por la red de distribución que los que aporta el propio nivel de presentación.

---

**Calidad de Experiencia.** La calidad de experiencia, mayormente conocida como QoE (Quality of Experience), por sus siglas en inglés, es la calidad percibida por el usuario final de nuestros servicios o aplicaciones. Esta percepción es subjetiva y se encuentra ligada a las expectativas que posea particularmente cada usuario. Veamos el enfoque que Batteram y otros mencionan sobre la QoE contrastada a la QoS:

se enfoca en los efectos percibidos por el usuario, como degradación en la calidad de la voz o el video, mientras que la QoS se enfoca en los efectos de la red, como demoras de extremo a extremo o jitter. (Batteram, y otros, 2010, pág. 176)

Si bien existe esta diferencia de enfoques, Montagud, Boronat, Martínez, Belda, & Cesar afirman que los parámetros de QoS “tienen un impacto muy claro sobre varios aspectos de la calidad de experiencia” (2015, pág. 1). Esto, situando lo dicho “En las comunicaciones telemáticas y, especialmente, en los servicios multimedia interactivos”. Estos servicios representan, justamente, nuestro objeto de estudio.

La QoE ciertamente se encuentra influida por la QoS, pero dos aplicaciones con similar calidad de servicios pueden presentar calidades de experiencia muy distintas. Podemos decir que, en este punto, dependerá de qué es lo que se hace con lo que se tiene, y no simplemente de qué es lo que se tiene. Si bien Montagud, Boronat, Martínez, Belda, & Cesar (2015) han centrado el objeto de análisis de su trabajo en los parámetros de la QoS que afectan la QoE, no significa que una dependa completa y únicamente de la otra. De hecho, si deseamos enfocarnos en brindar una experiencia de calidad, debemos enfocarnos en la experiencia desde un principio, y no solo en el servicio. Como indican Batteram, y otros, “Los proveedores de servicios se han centrado tradicionalmente en determinar y gestionar la QoS, no la QoE” (2010, pág. 177), y afirman que “este método no garantiza una estimación aceptable de la QoE para aplicaciones, sesiones o usuarios individuales”. De todas formas, también confirma la existencia de un vínculo entre ellas:

Por supuesto, la QoE está directamente relacionado con la QoS, pero el desafío para un proveedor de servicios es tener el conjunto correcto de herramientas y procesos para mapear la QoS a nivel de red a la QoE a nivel de usuario y sesión y tener la habilidad de controlarlo. (Batteram, y otros, 2010, pág. 176)

---

Esto no significa que Montagud, Boronat, Martínez, Belda, & Cesar (2015) estén equivocado o hayan olvidado algún aspecto, sino, que su objeto de estudio específico exigía otro punto de vista para el análisis. Como bien indican, “los niveles de Asincronía permisibles para los diferentes tipos de sincronización y tipos de datos multimedia (excepto para aromas) son bastante menores que las magnitudes de diferencias de retardos existentes en las redes actuales” (2015, pág. 6). Con esto fundamenta que, para brindar una buena QoE, con los niveles de QoS de la red de distribución, serán necesarias “soluciones precisas y adaptativas para la sincronización multimedia”. De hecho, como comentan Ibarrola, liberal, Taboada, & Ortega, “las herramientas de software para emular el comportamiento de la red son muy útiles, ya que permiten a los investigadores manipular fácilmente los parámetros de QoS para estudiar cómo las diferentes condiciones pueden afectar la percepción del usuario final.” (2009, pág. 1).

Como ya hemos indicado, la calidad de experiencia es un aspecto subjetivo. Adicionalmente, Ibarrola, liberal, Taboada, & Ortega concluyen en su investigación que “la aceptabilidad general de la aplicación (QoE) está estrechamente relacionada a otros parámetros contextuales, como las experiencias previas del usuario o sus expectativas.” (2009, pág. 4). De esta manera, la búsqueda de la mejor experiencia posible estará supeditada al conjunto de usuarios finales de la aplicación o servicio. Sabiendo que este será un grupo variable durante el tiempo, e incluso, que la subjetividad de un mismo usuario individual variará a medida que viva nuevas experiencias, tenemos una gran limitación en la búsqueda de la mejor calidad de experiencia posible. De acuerdo con Batteram, y otros, maximizar la QoE “También significa ser capaz de priorizar problemas, para abordar primero aquellos con el impacto más alto en términos de ingresos potenciales” (2010). Para esto, es importante tener indicadores adecuados, para obtener mediciones cuantitativas y actualizadas que podamos utilizar como fuente de toma de decisiones. También es importante tener en cuenta nuestro público objetivo, y trabar estas métricas con un grupo representativo del mismo. Por ejemplo, un director de orquestas podría tener menor tolerancia a los desfasajes de sincronización inter flujo que una persona que no suele escuchar ni crear música.

A la hora de elegir nuestros indicadores de calidad de experiencia, es importante tener en claro cuál es la experiencia que pretendemos brindar. Esto puede sonar un tanto obvio o redundante, pero no lo es. Por ejemplo, en el caso particular de este trabajo, no pretendemos una experiencia sonora con una sincronización exacta entre cada instrumento, sino, una experiencia

---

realista. Esto incluirá la posibilidad de desfasajes, en los casos en que los músicos estén tocando con dicho desfasaje. Se pretende una experiencia que refleje la realidad, sin añadir ni resolver problemas de sincronización. Por otro lado, se pretende una experiencia hacia el músico que brinde facilidades para no ejecutar de manera errónea las piezas. Si nuestros parámetros de medición apuntaran a la sincronización exacta entre todos los instrumentos, estaríamos tomando mediciones equivocadas. En este caso, quien ejecuta una obra debe tener también una experiencia de calidad que no lo lleve a cometer equivocaciones. Si la experiencia que recibe lo conduce a generar asincronías que no generaría en un conjunto presencial, habríamos fracasado en un objetivo importante de QoE.

### ***Umbral de Percepción Humana***

Se han realizado diversos estudios sobre la percepción humana de la sincronización. Los mismos, se encuentran enfocados en determinar qué rangos de sesgos son perceptibles por el ser humano, y cuales no lo son. “La diferencia temporal entre las unidades de datos lógicas de audio y video relacionadas es conocida como ‘sesgo’. Las transmisiones que están perfectamente ‘en sincronía’ no tienen sesgo” (Steinmetz, Human Perception of Jitter and Media Synchronization, 1996, pág. 61). En la actualidad, extendemos esta definición a la diferencia temporal que pueda haber entre las unidades de datos lógicas de dos flujos de medios, pudiendo ser estos de diversos tipos. Originalmente se limita la definición a un flujo de audio y otro de video, debido a que los estudios se enfocan en medir el sesgo entre estos. En principio, y a fines prácticos, mantendremos esta definición y nos centraremos en ello.

En el caso mencionado, la tolerancia es mayor cuando la imagen anteceda al sonido (sesgo negativo), y menor cuando la imagen precede al sonido, o el sonido antecede a la imagen (sesgo positivo). Este rasgo es apreciable ya desde 1993, en algunas, aunque no todas, de las pruebas de Steinmetz & Engler (Human Perception of Media Synchronization, 1993). También fue confirmado luego en los estudios de Bleumers, y otros (2012, pág. 454). Esto se justifica en el acostumbamiento humano a percibir imágenes antes que sonidos, dada la diferencia entre la velocidad de la luz y la del sonido. “Entonces, el sonido causado por un evento siempre, en realidad, llega a un observador más tarde que la luz de ese evento” (Salmon & Mason, 2008).<sup>2</sup>

---

<sup>2</sup> Se agradece especialmente los aportes de la investigación de “Impacto de Parámetros de QoS en Aspectos de QoE: Análisis desde el Punto de Vista de la Sincronización Multimedia” (Montagud, Boronat, Martínez, Belda, & Cesar, 2015) y sus bibliografías citadas para la confección de este apartado.

---

No todos los autores utilizan una misma forma de escribir un intervalo, algunos presentan primero el sesgo negativo y luego el positivo, y otros, al contrario. De todas formas, se suele apreciar normalizado el mencionar con signo negativo cuando la imagen se encuentra antes que el sonido, y con signo positivo cuando se encuentra después. Esto puede explicarse con la siguiente cita:

el flujo de audio siempre fue seleccionado como el flujo de referencia en el prototipo de sincronización maestro (audio) - esclavo (video). Fue elegido el audio debido a que la percepción humana es más sensitiva a las degradaciones de las señales de audio. (Huang & Nahrstedt, 2013)

Para estar seguros de que esta es la forma en la que deseó expresarse el autor, podemos valernos del fenómeno físico mencionado anteriormente y su efecto. El valor absoluto más grande, entre ambos del intervalo, corresponderá al video adelantado al audio, siempre que se hable de un rango de percepción humano. Por lo general, entonces, encontraremos intervalos con valores negativos mayores que los positivos.

En cuanto a la definición de un umbral de percepción humana, existen situaciones a tener en cuenta. Como puede evidenciarse en nuestro apartado sobre QoE, y de acuerdo a los estudios realizados por Ibarrola, liberal, Taboada, & Ortega, “predeciblemente, los umbrales de aceptabilidad cambian con el tiempo” (2009, pág. 4). Adicionalmente, podemos afirmar que “cuanto mejor es la resolución, más obvios se vuelven los errores de sincronización de labios” (Steinmetz & Engler, 1993). Posiblemente, no solo la resolución afecta el rango de percepción, sino la cantidad de cuadros por segundo. De todas maneras, es necesaria la realización de estudios más específicos para demostrarlo.

En 1993, los estudios de Steinmetz & Engler (Steinmetz & Engler, 1993) determinaron que un conjunto de audio y video con un sesgo de  $\pm 80$  ms se considera en sincronía, mientras que uno de  $\pm 160$  ms se considera fuera de sincronía. Un punto curioso de su estudio son las pruebas realizadas con profesionales de corte de video. Ellos afirmaron reconocer errores de sincronización de  $\pm 80$  ms, pero no creyeron que influyera en la calidad del contenido, e incluso uno de cada tres, reconoció errores de  $\pm 40$  ms. Lamentablemente, no ha especificado la resolución de la mayor parte de sus pruebas, las que fueron realizadas “en un monitor profesional

---

de alta resolución” (Steinmetz & Engler, 1993, pág. 3). Lo que podemos saber, es que se realizó a una resolución mayor a 240 x 256, ya que esta última la aclara como una resolución alternativa utilizada. Nuevamente Steinmetz, tres años más tarde, muestra resultados de una forma muy similar, aunque agrega que un experimento “con un sesgo de -240 ms o +160 ms condujo a una verdadera distracción del contenido y a un severo sentimiento de molestia” (Steinmetz, 1996).

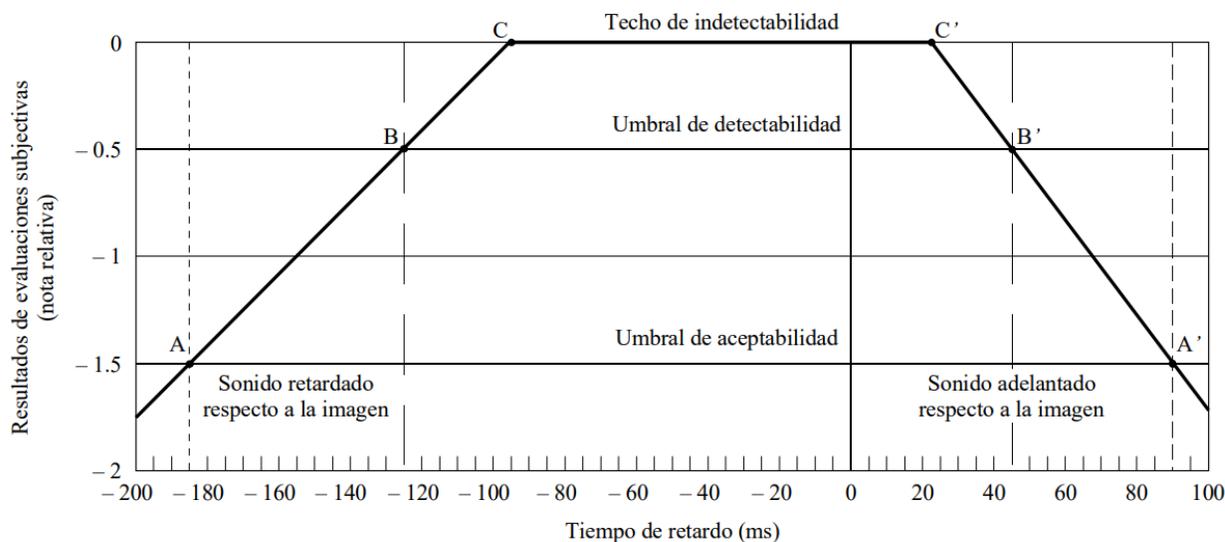
Varios años después, en 2003, Mued, Lines, Furnell, & Reynolds (2003), realizaron un trabajo similar, pero utilizaron una resolución menor, de 176 x 144. En consecuencia, la percepción fue menor, lo que nos indica que la resolución tiene mayor influencia que el paso del tiempo. Sus pruebas, fueron ejecutadas con sesgos de  $\pm 440$  ms, mayores que los mencionados hasta aquí. Aun así, en ninguna prueba alcanzó un nivel de aserción del 50%, de parte de los usuarios, sobre si el audio se encontraba antes que el video, viceversa, o se encontraba correctamente sincronizado. Esto demuestra que existe una gran dificultad en apreciar la desincronización con una definición de imagen tan pequeña. Consecuentemente, cuanto menor sea la resolución de un video, mayor será el sesgo permitido.

En 1998, se publicó la recomendación UIT-R BT.1359-1 (Unión Internacional de Telecomunicaciones, 1998). Como veremos, la misma continúa siendo una referencia válida hasta la actualidad. Esto revalida lo comentado en el párrafo precedente. El paso del tiempo tiene menos influencia que la calidad del video. Si se accede al sitio oficial de la Unión Internacional de Telecomunicaciones, se puede observar que esta recomendación sigue estando en vigor actualmente. Entre sus hallazgos, es de nuestro interés el siguiente fragmento:

las evaluaciones subjetivas muestran que los umbrales de detectabilidad son de unos +45 ms a -125 ms y que los umbrales de aceptabilidad son de unos +90 ms a -185 ms, como promedio, y que un valor positivo indica el adelanto de la señal de sonido respecto a la de imagen. (Unión Internacional de Telecomunicaciones, 1998, pág. 1)

**Figura 3.**

*Umbral de detectabilidad y aceptabilidad humana.*



*Nota.* Tomado de *Umbral de detectabilidad y de aceptabilidad* (pág. 4) por Unión Internacional de Telecomunicaciones, 1998, Rec. UIT-R BT.1359-1.

Staelens, y otros (Staelens, y otros, 2012), han realizado una revalidación de los datos presentados por tanto por la Rec. UIT-R BT.1359-1 (Unión Internacional de Telecomunicaciones, 1998), como por Human perception of jitter and media synchronization (Steinmetz, 1996), indicando haber obtenido resultados similares. Sus pruebas, se ejecutaron utilizando un monitor de 17 pulgadas a 1024 x 768 y, haciendo el cálculo correspondiente, asumimos que a una tasa de 25 cuadros por segundo. Indican que “retrasos de hasta un cuadro de video [-40, 40 ms] no se detectan en absoluto” (Staelens, y otros, 2012, pág. 454). También obtuvieron que un sesgo de -240 ms fue percibido por todos los sujetos de prueba, y en uno de 120 ms (sesgo positivo), solo un tercio de los participantes logro detectarlo. Adicionalmente, indican que “retrasos de hasta 2 cuadros de video [-80, 80 ms] solo son detectados por una pequeña cantidad de sujetos” (Staelens, y otros, 2012, pág. 454). Dichos autores, tal como Steinmetz & Engler (1993), también realizaron un contraste de pruebas entre usuarios expertos y no expertos. En este caso, los usuarios expertos eran traductores, cuya tarea se basaba en oír y traducir en vivo el contenido. Los resultados de los traductores no fueron mejores que los de los

---

usuarios no expertos, presuntamente, por haber estado más concentrados en la propia tarea de traducción que en el análisis de sincronización.

El trabajo más reciente que encontramos dentro del campo fue realizado en 2020, se trata del Trabajo de Fin de Grado titulado “Detección de Sincronización Audiovisual Utilizando Deep Learning” (Román Sarmiento, 2020). En este trabajo, la autora se nutre de los valores de la Rec. UIT-R BT.1359-1, lo cual consideramos correcto, siendo un organismo internacional que afirma que la misma se encuentra en vigor. Montagud, Boronat, Martínez, Belda, & Cesar (2015) también han citado la fuente, pero sin emitir un juicio de validez sobre ella, si no, presentándola entre otras fuentes.

El Advanced Television Systems Committee establece que, para la televisión digital, “El programa sonoro nunca debe adelantarse al programa de video por más de 15 milisegundos, y nunca debe retrasarse al programa de vídeo en más de 45 milisegundos” (2003, pág. 4). Pese a que luego se le añade una tolerancia extra de  $\pm 15$  ms en el lado del receptor, para presentar el contenido, se encuentran entre los valores más bajos mencionados hasta aquí. En dicho documento no se presenta como fundamento ningún análisis de datos. Igualmente, en comparativa, podemos observar que apuntan al rango de imperceptibilidad absoluta. De esta manera, los valores mantienen consistencia con los mencionados hasta aquí.

Algunos de los estudios mencionados fueron realizados específicamente analizando la sincronización de labios. Estos umbrales de percepción resultan perfectamente aplicables a nuestro objeto de estudio, ya que, de acuerdo con Steinmetz, “La presentación de un violinista en un concierto, así como un coro, no mostraron demandas de sesgos más estrictas que un orador” (1996, pág. 63). También debemos mencionar que la sincronización audio/video no es la única en la que entra en juego la percepción humana. De hecho, los rangos de percepción para relaciones de audio/audio resultan mucho más restrictivos, al menos para audios acoplados. Su impacto en la percepción dependerá también de cuán acoplados se encuentran los flujos de medios que presentan el sesgo.

### ***Umbral de Error Humano***

Laguna & Shifres (2012) realizaron pruebas en las que miden la sincronización de acompañantes musicales y bailarines, mientras realizan obras de manera conjunta. Si bien el estudio no se basa en la sincronización entre dos músicos, es una buena muestra del umbral de error humano, ya que el músico y el bailarín van ajustando sus interpretaciones para ejecutarlas

---

de manera sincronizada. Adicionalmente, y lo que más nos interesa, miden los desvíos del acompañante musical contra un pulso sonoro a velocidad constante.

En este caso, para comprender el análisis, debemos considerar que se encuentra basado en una velocidad de negra de 60,14. Esto significa, 60,14 golpes por minuto, o 1 golpe cada 997,6720984369804 ms. Los desvíos que presentan se expresan en golpes, con lo cual, un desvío de 1 golpe correspondería a la cantidad de ms recién mencionada, y uno de 0.01 golpes, a 9,976720984369804 ms. Para ser más prácticos, de aquí en adelante diremos que representa aproximadamente 10 ms. También es preciso considerar que, al realizar una tarea en conjunto con un bailarín, existen períodos donde, al detectar asincronías, ambos comienzan a ajustar sus interpretaciones con el objetivo de realizarlas de manera sincronizada. Esto es mencionado en la bibliografía como la *negociación* entre el bailarín y el acompañante musical.

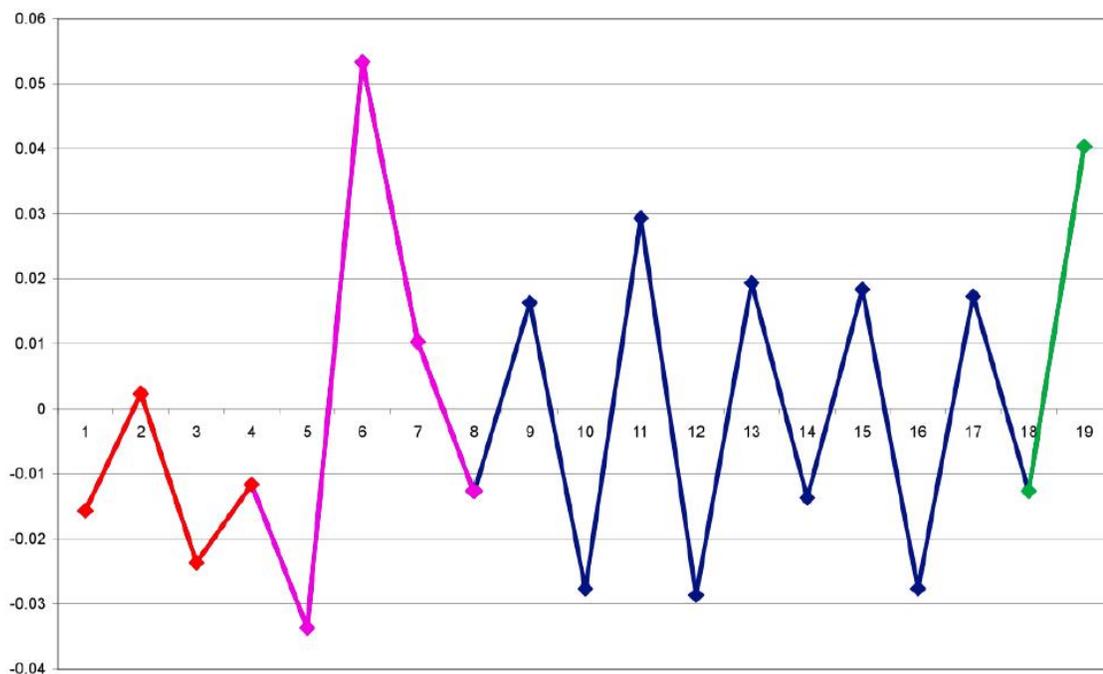
Resulta de nuestro interés la variación de desvíos obtenida entre golpes ejecutados por un acompañante musical y el momento en el cual debió ejecutarse cada golpe, representado por el pulso sonoro. Si bien es cierto que se encuentra condicionado a la negociación mencionada, esto justificaría aumentos o decrementos de desvío levemente progresivos entre golpes. Por contrario, se puede observar que, mayormente, cada golpe ejecutado tuvo un desvío inverso al anterior, es decir, si un golpe se adelantó a lo esperado, el siguiente se postergó a lo esperado, y viceversa. Esto puede atribuirse a un intento de marcado de los golpes fuertes, como indica el autor “el patrón cambia sensiblemente, a partir del compás 3 (en azul) haciéndose más notable la diferencia entre los beats impares (fuertes) y pares (débiles), excediendo claramente el umbral de detectabilidad mencionado” (Laguna & Shifres, 2012, págs. 9-10). Anteriormente, Laguna & Shifres, mencionan que los golpes del primer compás (primeros 4 golpes, marcados en rojo) se encuentran dentro del umbral (2012, págs. 8-9).

Para tener una dimensión más clara del umbral de error humano que se presenta, se puede analizar la Figura 4. Como se puede apreciar, existen variaciones leves y otras más marcadas, siendo algunas perceptibles y otras no, de acuerdo con el umbral humano. Visualmente, y teniendo en cuenta la proporción de aproximadamente 10 ms cada 0.01 golpes, se observa que las mayores desviaciones son de algo más que -30ms y 50ms, mientras que la más pequeña no llegaría a los 5 ms. Cabe destacar que, al tener una variación negativa seguida de una positiva, la diferencia de sus desvíos habría sido de más de 80ms. Es decir, si se esperara percibir un golpe

aproximadamente 998 ms después que el anterior, en lugar de compararlos con un pulso constante, el desvío entre uno y otro tendría dicho valor.

#### Figura 4.

*Desfasaje en golpes ejecutados por un acompañante musical de danza.*



*Nota.* Tomado de *Perfil de timing de la ejecución del AMD en el estímulo en términos de diferencia de duración del IOI (inter-onset-interval) real respecto del nominal* (pág. 8) por Laguna & Shifres, 2012, *Indicios Visuales y Auditivos en el Ajuste Sincrónico del Pulso Subyacente Entre Bailarines y Acompañantes Musicales.*

Malbrán (2007) realizó un estudio con una muestra de 40 sujetos, compuesta por “Músicos con título terciario en música” (2007, pág. 4) y “Estudiantes universitarios de nivel avanzado (tercer año de la carrera).” (2007, pág. 4). En su estudio, la medición cuantitativa no se dividió en ms sino en tics. “Cada ataque (tactus) se dividió en 480 tics” (2007, pág. 5). De esta forma, el margen de error de sus mediciones se corresponde al valor en ms de 1 tic, lo cual es un margen bastante restringido. Como veremos, la duración de 1 tic no supera, en ningún caso, 1,14 ms. En su estudio se utilizaron 3 variantes de tiempo de negra, 110, 126 y 141 negras por minuto.

A continuación, expresamos la duración de cada negra y cada tic en ms, ya que es la unidad de medida que utilizamos en nuestro trabajo.

**Tabla 1.**

*Tiempos de ejecución de TIC expresados en milisegundos.*

Negras por minuto	seg por negra	ms por negra	ms por TIC
110	0,545454545	545,4545455	1,136363636
126	0,476190476	476,1904762	0,992063492
141	0,425531915	425,5319149	0,886524823

*Nota.* Esta tabla fue confeccionada para comprender la duración de los tics que utilizó Malbrán (2007) en su estudio. No representan una unidad de tiempo estándar.

Estos valores de ms por tic nos son de suma importancia para comprender la Tabla 2. En ella, se presentan las cantidades totales de golpes analizados, y cuántos y que porcentaje de ellos fueron en tiempo, adelantados, atrasados y fuera de la ventana definida. Ello, teniendo en cuenta los datos que se presentan en la siguiente cita.

El punto de ataque para cada tactus se consignó en tics.

- La sincronía estricta es el punto cero.
- La ventana de tolerancia se corresponde con el software en el rango 60 a +60.

Los ataques en sincronía negativa (adelantados) van de -1 a -60.

- Los ataques en sincronía positiva (atrasados) van de 1 a + 60.
- Los ataques retrasados que exceden la ventana de tolerancia (off ventana) van del tic 61 a 420 (Malbrán, 2007, pág. 5)

**Tabla 2.**

*Golpes certeros, adelantados o atrasados, dentro y fuera de la ventana de tolerancia.*

Tempo	Ventana de tolerancia									Off vent.	
	N°tot..	cero	%	Adel.	%	Atrás.	%	Tot vent	%		%
110	3089	46	1,48	781	25	2068	67	2895	94	194	6
126	2899	30	1,03	484	17	2053	71	2567	89	332	11
141	3489	19	0,54	234	7	2192	63	2445	70	1044	30

*Nota.* Tomado de *Tabla 1* (pág. 5) por Malbrán, 2007, *Sincronía Rítmica y Tempo: Un Estudio con Adultos Músicos*.

Como se observa, el porcentaje de golpes realmente certeros es muy acotado, en ningún caso llega al 1,5%. La mayoría de los golpes se encuentran, en los tres casos, atrasados, pero dentro de la ventana prevista. A continuación, se presentan los intervalos de tiempo, expresados en ms, en los cuales rige esta medición. Con ello podemos comprender que el resultado es lógico, ya que el punto de sincronía cero es de aproximadamente  $\pm 0,57$  ms,  $\pm 0,50$  ms y  $\pm 0,44$  ms, y los intervalos de adelanto y atraso se encuentran más relacionados con los umbrales de percepción humana.

**Tabla 3.**

*Rango de milisegundos considerado adelantado, cero, atrasado o fuera de rango.*

Tempo	Adelantados (-1 a -60)	Cero	Atrasados (1 a 60)	Fuera (61 a 420)
110	(-68,749999978; -0,568181818) ms	(-0,568181818; 0,568181818) ms	(0,568181818; 68,749999978) ms	(68,749999978; 477,840908938) ms
126	(-60,019841266; -0,496031746) ms	(-0,496031746; 0,496031746) ms	(0,496031746; 60,019841266) ms	(60,019841266; 417,162698386) ms
141	(-53,6347517915; -0,4432624115) ms	(-0,4432624115; 0,4432624115) ms	(0,4432624115; 53,6347517915) ms	(53,6347517915; 372,7836880715) ms

*Nota.* Esta tabla fue confeccionada para comprender, expresando en milisegundos, el rango de desfasajes a los que Malbrán (2007) llama, en su estudio, adelantados, cero, atrasados o fuera de rango.

Luego, Malbrán (2007, pág. 6) presenta una tabla con los valores mínimos, máximo, medios y el desvío estándar sobre esta muestra. La misma, la presentaremos adosando los correspondientes valores en ms para nuestro análisis, y conservando los originales expresados en tic por si fuesen de interés al lector.

**Tabla 4.**

*Desvíos mínimos, máximos, medios y desvíos estándares en la ejecución musical.*

Negras por minuto	seg por negra	ms por negra	ms por TIC	TIC				ms			
				Mínimo	Máximo	Media	Desvío estándar	Mínimo	Máximo	Media	Desvío estándar
110	0,54545455	545,454546	1,13636364	-67	419	23,47	53,28	-76,1363636	476,136364	26,6704546	60,5454546
126	0,47619048	476,190476	0,99206349	-50	406	28,91	42,31	-49,603175	402,777778	28,6805556	41,9742064
141	0,42553192	425,531915	0,88652482	-55	414	55,31	60,23	-48,7588653	367,021277	49,0336879	53,3953901

*Nota.* Adaptado de *Tabla 2* (pág. 6) por Malbrán, 2007, *Sincronía Rítmica y Tempo: Un Estudio con Adultos Músicos*.

Como podemos observar, al menos para los tiempos de negra 110 y 126, la media de ejecución de los golpes se encuentra atrasada por debajo de los 30 ms. Esto es un valor aceptable teniendo en cuenta el rango de percepción humana. Para el tiempo de negra 141, alcanza los 49 ms, lo cual deja de ser tan aceptable. De todas formas, si tenemos en cuenta los valores máximos, mínimos y el desvío estándar, obtenemos desvíos que superan en todos los casos los  $\pm 40$  ms. En cuanto a los máximo y mínimos, que alcanzan los -76 ms y 476 ms, debemos tener en cuenta que se trata de la peor ejecución de miles de golpes tomados como muestra. No obstante, existen, y demuestran que el error humano puede superar ampliamente la percepción del mismo. Los resultados obtenidos son muy variables entre sí y, como indica la autora, “4 Investigaciones previas coinciden en señalar la alta variabilidad de las respuestas de sincronía con patrones isócronos y difieren en las técnicas utilizadas para estimar dicha variabilidad” (Malbrán, 2007). El desvío estándar puede considerarse como la ejecución que esperamos normalmente de un músico, la cual dista de ser completamente sincronizada, y se encuentra dentro del rango de detectabilidad, mayormente en el caso de negra 110, donde superando los 60 ms.

Shifres & Holguín Tovar (2015) nos muestran que, si bien algunas habilidades musicales auditivas y cognitivas se encuentran más desarrolladas en algunas personas que en otras, aún sin

---

haberlas ejercitado antes, también pueden entrenarse. Podemos valernos de sus dichos al hablar de la capacidad que tiene una persona para ejecutar de manera más o menos sincronizada las obras musicales, dado que dicha capacidad depende de una habilidad auditiva y cognitiva. Una de las áreas de desarrollo musical del panorama actual de desarrollo de habilidades auditivas es, como indican Shifres & Holguín Tovar, “ajuste temporal y tonal en las ejecuciones tanto vocales como instrumentales” (Shifres & Holguín Tovar, 2015, pág. 13). De esta forma, se confirma que el ajuste temporal puede mejorar ejercitándolo. Para afirmar que un músico posee este tipo de habilidades desarrolladas, también podemos apoyarnos en los requisitos que debe tener para ejecutar buenas imitaciones. Dentro de ellos, se encuentran habilidades relacionadas con la precisión y estabilidad temporal.

Esta habilidad se apoya en un conjunto de requisitos que el sujeto debe disponer para efectuar una imitación deseable, entre los que podemos destacar la afinación, la precisión rítmica y métrica, el conocimiento de la organización del discurso, la anticipación, el mantenimiento de un tempo estable, las nociones de tónica y dominante, entre otras. (Shifres & Holguín Tovar, 2015, pág. 134)

De acuerdo con esto, Malbrán expresa que “Los valores de anticipación varían según el grado de entrenamiento” (2007, pág. 6), luego de afirmar “La respuesta para ser sincrónica requiere un grado calculado de anticipación” (Malbrán, 2007). Esta anticipación se basa en que una persona no puede reproducir un sonido en sincronía con otro por el simple hecho de oírlo. Es necesario que, en base a su conocimiento del tiempo a seguir, estructure los golpes en su mente para anticiparse a ellos, “si espera la aparición de cada ataque llega tarde” (Malbrán, 2007, pág. 3). En base a esta anticipación, el individuo ejecuta luego una serie de acciones motoras.

Las acciones de sincronización permiten estudiar cómo las personas codifican la duración, esto es el modo en que representan en su mente el intervalo que separa los estímulos con los que deben sincronizar el gesto motor. Sincronizar es traducir el ritmo perceptivo a ejecución motora. (Malbrán, 2007, pág. 2)

Con lo dicho hasta aquí, afirmamos entonces que una persona que practica y/o estudia para ello, tenderá a reducir su margen de desvíos al ejecutar sonidos sincronizados. Sin embargo,

---

esto no garantiza que su umbral vaya a ser más ajustado que el de otra persona que no lo ejercita. La mejoría, cuando se produce, lo hace sobre el propio umbral de quien práctica, el cual podría partir de una habilidad perceptiva menor que la de otra persona. “Tener un buen oído musical es una capacidad ampliamente demandada y concomitantemente valorada en los músicos” (Shifres & Holguín Tovar, 2015, pág. 11). No todas las personas lograrán obtenerlo en la misma medida e, incluso, “son numerosas las personas que se enfrentan a una gran frustración por no alcanzar el desarrollo deseado para brindar, a partir de la audición musical, el tipo de respuestas estructurales esperadas” (Shifres & Holguín Tovar, 2015, págs. 27-28).

El margen de desvío en la ejecución musical se encuentra directamente relacionado con el umbral de percepción humana, con su habilidad para predecir y con su habilidad motriz para ejecutar dicha previsión. Incluyendo esto, el umbral de percepción se concluye que el umbral de error de una persona no será menor al de percepción que la misma posea. De hecho, será un poco mayor, teniendo en cuenta el tiempo que demore en ejecutar la acción que desencadene, lo que Malbrán nombra como “timing perceptivo-motor” (2007).

### ***Margen Tecnológico***

Con este apartado pretendemos concientizar sobre la existencia del retraso agregado, tanto en la captura como en la muestra de sonido e imágenes, por los distintos equipos tecnológicos que se conectan a un computador para ello. Como se ha mencionado en el apartado de Calidad de Servicio, de la página 24, existen también tiempos de ejecución en procesador. Estos tiempos resultan medibles gracias al reloj del equipo. En cuanto a la captura y la presentación, no solo distinto hardware, sino distintas condiciones físicas, específicamente la distancia, pueden afectar al retraso total. Sencillamente, el audio es capturado por un micrófono y entregado por un parlante, pero no todos los micrófonos y/o parlantes son idénticos entre ellos, ni todas las personas los ubican de la misma forma, o en el mismo lugar. Esto sucede también con la imagen, capturada por una cámara y entregada por una pantalla.

En cuanto a los micrófonos, podemos tener en cuenta la tabla que presentan Salmon & Mason (2008), donde muestran los milisegundos que tarda en capturar el audio cada tipo de micrófono, basándose en la distancia entre micrófono y la boca del hablante. A grandes rasgos, podemos notar una diferencia significativa entre un micrófono ubicado a 7 metros y otro ubicado a 2 cm de los labios. La misma es presentada a continuación como la Tabla 5 de nuestro trabajo.

**Tabla 5.**

*Demora aproximada agregada por distintos micrófonos.*

<b>Mic type</b>	<b>Distance</b>	<b>Delay</b>
lip mic	2cm	~0.1ms
tie-clip/lavalier mic	30cm	~1ms
desk/stick mic	60cm	~2ms
boom mic	1m - 2m	~3ms - 10ms
shot-gun mic	1m - 7m	~3ms - 20ms
camera-mounted mic	1m - 7m	~3ms - 20ms

*Nota.* Tomado de Salmon & Mason, 2008 (pág. 7), Factors affecting perception of audio-video synchronisation in television.

En contraposición, también presenta los valores aproximados que demora el sonido en llegar desde los altavoces hacia nuestros oídos en distintos ámbitos, basándose también en las distancias físicas. Dichos valores son presentados a continuación como Tabla 6 de nuestro trabajo.

**Tabla 6.**

*Demora aproximada agregada por distintos altavoces, en función a la distancia.*

cinema	~50ms
home cinema	~15ms
living room	~10ms
bedroom	~5ms
handheld	~2ms
headphones	~0ms

*Nota.* Tomado de Salmon & Mason, 2008 (pág. 7), Factors affecting perception of audio-video synchronisation in television.

Una vez que el audio del habla llega hacia el micrófono, podríamos creer que la latencia agregada tiende a 0 ms, al menos si el micrófono y el dispositivo se conectan por medio de un

---

cable, no obstante, podría haber casos en que no sea así. Eduardo Quintero, Technical Director Latin America y Sales & Marketing Director Mexico & Central America, en Audio-Technica América Latina S.A., nos ha respondido un correo al cuál consultamos sobre cuántos milisegundos de retraso agregarían sus equipos. De acuerdo con Quintero (Comunicación personal, 2022), un micrófono tendría cero latencias al entregar la señal capturada, aunque podrían agregarse algunos milisegundos dependiendo, por ejemplo, de la calidad del convertidor de audio analógico a digital que posea. De igual manera, Mark E. (Comunicación personal, 2022) nos ha mencionado al DAC (convertidor digital a analógico), junto a los códecs utilizados, como algunos motivos que pueden afectar estos tiempos en el caso de los parlantes. Mark E. es el Customer Support Manager de Audeze, quién también accedió a brindar esta información por correo electrónico. Salmon & Mason (2008) reconocen la existencia de este pasaje de señal como una demora, aunque la cataloga como insignificante teniendo en cuenta los rangos de percepción humanos. Es claro que un desvío de unos pocos milisegundos no producirá una asincronía perceptible, pero debe tenerse en cuenta como parte de la sumatoria de varios desvíos.

La señal eléctrica producida por el micrófono puede ser analógica o digital. La señal representa, de forma sencilla, la presión del aire, o gradiente de presión, en el micrófono.

La producción de señales digitales implica una frecuencia de muestreo y un proceso de muestreo y cuantificación. Para el audio, estos procesos tienen un efecto insignificante sobre la percepción de la sincronización. (Salmon & Mason, 2008, pág. 7)

También debemos considerar la imagen emitida por una pantalla, o capturada por una cámara. Centrándonos en las pantallas, existen distintos tipos de ellas. Mencionaremos, a grandes rasgos, monitores, televisores y proyectores. Dentro de ellos existen distintas tecnologías, sea en el panel o el tipo de proyección. Si bien la mayoría de los modelos, y de diversos fabricantes, no brindan información sobre la latencia públicamente, sí lo hacen aquellos que enfocan su público objetivo en los videojuegos. Esto, dado a la búsqueda de la menor latencia posible dentro de dicha actividad. La falta de datos en otros equipos posiblemente significa que los valores de este segmento son realmente menores a los de pantallas convencionales. Aun así, dentro de aquellos que se ofrecen como específicamente para videojuegos, podemos encontrar diferencias de, al menos, 18 ms entre uno y otro. Por ejemplo, el AORUS FI27Q-X Gaming Monitor, como se

---

observa en el sitio oficial de Aorus (2022), presenta dos modos, uno de los cuáles ofrece una latencia de 0.3 ms y otro de 1 ms. En cambio, el Televisor LG 4K UHD (modelo 55UQ81006LB), presenta una latencia menor a 19 ms, según informa LG (2022). El Proyector BenQ TK700STi, promocionado como “World’s First 4K HDR Gaming Short Throw Projector” (BenQ, 2022), presenta distintas latencias dependiendo su configuración, siendo 16.7 ms (1080p@60Hz o 4K@60Hz), 4.2 ms (1080p@240Hz) y 8.3 ms (1080p@120Hz). La distancia física entre nuestros ojos y la pantalla también produce un tiempo adicional, el cuál será considerablemente menor al mencionado para el sonido. En la investigación de Salmon & Mason (2008) podemos observar también que un televisor adiciona una demora al recibir una cantidad de imágenes por segundo y estar configurado, o diseñado, para mostrar una cantidad diferente. Esto también se presenta, y con valores influyentes, en cuanto a la definición de la imagen.

El valor medio de los resultados de la prueba para las imágenes en definición estándar fue 18 ms mayor que para las imágenes de alta definición. Por medición objetiva (usando un osciloscopio y lápiz óptico) se estableció que el televisor LCD introdujo un retardo extra de video de 20 ms trabajando en definición estándar. (Salmon & Mason, 2008, pág. 14)

### ***Coros y Orquestas***

Un aspecto interesante que puede encontrarse en el trabajo de Laguna & Shifres (2012) es la negociación de la cual hablan, que implica que entre el bailarín y el instrumentista ajustan lentamente, tiempo a tiempo, la posible desincronización que perciben en el momento. Consideramos que cuando se debe actuar sobre una base inmodificable, se pierde esta posibilidad, produciendo que en lugar de acercarse el uno al otro, el ajuste temporal deba ser solo de parte de uno. De todas maneras, en un conjunto instrumental de más de 2 personas, esta negociación es un poco más compleja y no se ve tan facilitada como en un dúo.

Al hablar de coros y orquestas, estamos ciertamente hablando de grupos con una mayor cantidad de integrantes, de aquí, la necesidad de tener un conciliador temporal para las ejecuciones de las obras. Actualmente, al hablar de grupos grandes y completos, quien ocupa este lugar es el director, pero no siempre fue de esta manera. Tampoco sería justo definir a un director como un conciliador temporal, ya que sus funciones son mucho más abarcadoras.

---

En su análisis de los orígenes, desarrollo y evolución de la dirección de orquesta, Villarreal Rodríguez (2016, pág. 5) menciona al, una vez llamado, jefe de la orquesta, conocido hoy como el concertino. El mismo, realizaba una tarea de dirección dividida junto con el propio compositor de las obras. Este rol no deja de ser ocupado por un instrumentista, que suma a su tarea de ejecución musical la de guiar al resto de instrumentistas con respecto al tiempo. “representado regularmente por un violinista (lo que hoy conocemos como concertino) o un clavecinista” (Villarreal Rodríguez, 2016, pág. 5). Cabe mencionar que aún en la actualidad, en grupos más pequeños o informales, puede no existir la figura del director, o bien existir, pero por necesidad del conjunto ocupar también un rol de instrumentista.

**La Figura del Director.** Como hemos mencionado, la figura del director es más abarcadora que únicamente ser un conciliador temporal, adicionalmente, debemos mencionar que sus tareas comienzan antes de la presentación musical ante un público.

La dirección musical —ejemplificada aquí en el estándar de la dirección orquestal— comprende tres procesos: el estudio individual, el trabajo de ensayo y la ejecución en el concierto. Probablemente, de ellos es el más relevante para el hecho musical el trabajo de ensayo, ya que es en éste donde se condiciona en gran medida el resultado musical. (Martínez Pegalajar, 2019, pág. 179).

En cuanto a las tres tareas mencionadas, las mismas se coinciden con los dichos del director de la orquesta filarmónica de Bogotá, Francisco Retting, quien remarca la importancia del estudio individual, al ser entrevistado por Molinares, Ione. “Un director de orquesta nunca debe llegar ante la orquesta en un estado de no preparación” (Retting, 1990). Según menciona, primero debe estudiar la obra y, entendiendo cada detalle de la misma, tener en su mente todos los sonidos como el autor quiso expresarlos, luego de ello, puede comenzar el ensayo con la orquesta. Entre estas tareas, para Martínez Pegalajar, “debido a la situación central y a la relevancia del director, es inevitable entender cuál es su función dentro del ensayo de conjunto” (2019, pág. 179).

Dentro del ensayo es donde el director transmitirá a los integrantes de la orquesta todo lo estudiado. Es claro que cada uno de los integrantes pueden y deben haber estudiado por su cuenta, y antes del ensayo, las partituras de las obras. Esto significa que tendrán en su mente

---

también una forma en la que debe sonar, al menos, su propio instrumento. Posiblemente, haya diferencias en la concepción de un músico y el director, o entre dos músicos diferentes, subgrupos diferentes, etc. “la diversidad de opiniones y soluciones lleva a la necesidad de depositar la responsabilidad en una sola persona, el director” (Martínez Pegalajar, 2019, pág. 189), surgiendo así su tarea de concertación.

La concertación es una acción que viene del término concertar, definido por Pena, Anglés, & Querol como “ajustar o concordar a los ejecutantes de una obra” (Pena, Anglés, & Querol, 1954, pág. 555). Esta misma definición es utilizada por González Lapuente (2003, pág. 136). En otras palabras, el director de orquesta debe poner de acuerdo a las distintas partes sobre cómo debe realizarse la obra en todos sus aspectos. Esto bien es dicho en pocas palabras, sin embargo, Previtali (1969) lo define mencionando de forma más detallada las labores que llevan a esa concordancia conjunta y, adicionalmente, menciona la tarea posterior que se realiza para que la ejecución de la obra concuerde también con lo ya convenido.<sup>3</sup>

La concertación es la labor que el director realiza durante los ensayos para explicar la partitura y lograr su traducción sonora en la ejecución instrumental y vocal; labor a la que se suma también el trabajo de entrenamiento para superar las dificultades técnicas. (Previtali, 1969, pág. 30)

Dentro de las dificultades técnicas, o problemas que deben ser detectados, también por el director durante el ensayo para su concertación, cabe destacar que el director deberá tener la capacidad de priorizarlos. El tiempo de un ensayo y la cantidad de ellos previos a una obra, como en todo proyecto con fecha de finalización, es limitado, por lo cual es posible que no puedan resolverse todos los problemas. Martínez Pegalajar (2019) detalla, en su trabajo, una clasificación de los mismos para su posterior priorización. También puede consultarse el mismo si se desea mayor información sobre las técnicas de ensayo que menciona en la siguiente conclusión. La misma es de nuestro interés para seguir abarcando las funciones del director.

---

<sup>3</sup> Se agradece especialmente los aportes de la investigación de “El director de orquesta en el ensayo: análisis teórico y práctico” (Martínez Pegalajar, 2019) y sus bibliografías citadas para la confección de este apartado.

---

Conviene entonces entender el ensayo desde dos perspectivas complementarias: como formato —cuya organización temporal, logística y musical implica lo que definiremos como técnicas de ensayo— y como labor de concertación, entendida como el trabajo musical específico del director durante el tiempo concreto en el que la actividad tiene lugar. (Martínez Pegalajar, 2019)

Dicho todo esto, podemos fundar nuestra afirmación, de la página 45, sobre lo abarcadoras que son las tareas del directo. Esto involucra el estudio de las obras, organización del ensayo y técnicas de ensayo, transmisión a la orquesta de lo estudiado, concertación de y con las partes, incluyendo qué y cómo corregir, trabajo de entrenamiento en cuestiones técnicas individuales o de conjunto, priorización de problemas, memorización de los que no se abarcan inmediatamente, y guiar en la transmisión del mensaje durante la ejecución de la obra. Todo esto, sin olvidar los aspectos psicológicos del grupo de personas con el cuál trabaja, las situaciones que pueden afectarlos y las cuestiones humanísticas a tener en cuenta en base a ello.

“La paciencia, el respeto y el trato con integridad, el sentido del humor, la capacidad de escucha, la calma, la fortaleza son cualidades psicológicas que constituyen el carisma que tanto necesita el director” (Chuang, 2005, págs. 17-18). Lo cual indica que “unido a una cierta capacidad de comunicación, le permite cumplir, como líder, como pedagogo y como músico, con sus funciones psicológicas dentro de la agrupación”. Retting (1990) menciona que, en el trabajo de director, se debe enfrentar el problema de las relaciones humanas, ya que recibe los problemas, en ocasiones, de 100 personas, y debe tener un sentido humano muy amplio, con una receptividad y tolerancia que llega a comparar con las de un matrimonio.

**Utilidad del metrónomo.** Es correcto mencionar que el metrónomo es un elemento u objeto, no obstante, se encuentra ligado en cierta forma a la figura del director de orquesta en sus orígenes. “La primera gran figura de la dirección de orquesta fue sin duda Jean Baptiste Lully (1632-1687)” (Villarreal Rodríguez, 2016, pág. 5). Como el autor comenta, Lully, para guiar a los distintos músicos de su orquesta, tomaba un bastón o una vara larga y comenzaba a golpear con él el suelo, de una forma lo suficientemente fuerte para que toda la orquesta lo oiga mientras se ejecutaban las piezas musicales, de esta forma indicaba el compás. “La importancia de Lully

---

reside en el aporte, en una versión primitiva de lo que hoy conocemos como metrónomo” (Villarreal Rodríguez, 2016, pág. 6).

El metrónomo fue construido por Johann Nepomuk Mälzel en 1816, a partir de una invención original de un tal Winkel en Amsterdam, por lo cual ocurriría con posterioridad un extenso pleito sobre la patente del aparato, pleito que sería ganado por este último en un tribunal holandés.

El metrónomo consta de un péndulo que oscila sobre un pivote con la ayuda de un mecanismo de reloj. Una especie de tic-tac señala la frecuencia de las oscilaciones y con la ayuda de un peso deslizante es posible controlar el número de oscilaciones por minuto. (Valencia Restrepo, 2001, pág. 2)

Actualmente, este artefacto puede encontrarse en versiones mecánicas, digitales y hasta en softwares para diversos dispositivos, los cuales también son regulables y emiten los pitidos por medio de parlantes. “El pulso es el marco de referencia básico, el latido fundamental de la música. La función del metrónomo es indicar el pulso” (Popoca, 2016, pág. 77). Su utilidad y aceptación parecen ser indiscutibles, aparentemente, seguir su ritmo presenta mayor facilidad que la de seguir el ritmo de una persona, al menos por oído. Como indica Malbrán, “El papel de la anticipación varía según se trate de acoplar la acción a una fuente de estimulación mecánica o expresiva. Pareciera ser más fácil sincronizar con ejecuciones mecánicas sintetizadas por computadora que con ejecuciones naturales” (2007, pág. 3). Teniendo en cuenta que, como observamos en la página 46, el trabajo de ensayo sería el proceso más relevante entre las tareas de un director, podemos apreciar cuán aceptada ha sido la ayuda de los pulsos mecanizados en el siguiente fragmento.

Sería muy saludable también la invención de un director-autómata, especie de tablero eléctrico, colocado en cada atril, con capacidad de indicar el tiempo y la intensidad como existe el conductor-automático para los aviones. El director de orquesta podría concentrar su atención en el trabajo intelectual, durante los ensayos. (Steinberg, 1946, pág. 17)

---

## **Contribuciones Principales**

No existen diferencias absolutas entre la percepción de sincronización de un humano y la capacidad de ejecución musical sincronizada de otro.

La utilización de la marca temporal de las pistas base, replicada en las pistas ejecutadas por los músicos, permite una consecuente sincronización inter paquete (Inter-Bundle).

Al realizar una reproducción y grabación simultaneas, en un equipo, dependiendo del mismo reloj, se minimizan los efectos contrarios que podría ocasionar la deriva del reloj local.

La gestión de solicitudes de paquetes en tiempo real, ejercida por el servidor, reduce los efectos negativos que pueda tener un bajón de flujo de bits de ellos.

Adicionando un tiempo de retraso de extremo a extremo que resulta despreciable para la calidad de experiencia de los consumidores finales, es posible realizar transmisiones multiusuario sincronizadas de obras musicalizadas.

## **Estructura General de la Tesis**

### ***Capítulo 1 – Introducción y Naturaleza del Problema***

El primer capítulo de este trabajo nos adentra en el problema alrededor del cual se basa la investigación realizada, brindando, además, todos los elementos teóricos necesarios para abordarlo. Podemos encontrar en él una subdivisión natural de tres etapas. En principio, la propia presentación del problema, que se compone tácitamente por su identificación, planteamiento, justificación. En segundo lugar, y en base al problema presentado, se presentan los objetivos de este trabajo, nuestra hipótesis, las variables sobre las cuáles se comienza la investigación y la propuesta de resolución, teniendo en cuenta que, dicha propuesta, no deja de ser un elemento introductorio que será ampliado en los capítulos posteriores. Por último, se elabora el marco teórico que rodea a nuestro problema. Para esto no solo se tienen en cuenta aspectos tecnológicos, sino, los aspectos que inevitablemente deben comprenderse para poder analizar por completo nuestro problema. Más específicamente, musicales y perceptivos. La percepción es un elemento que debe considerarse ligado a la sincronización de archivos y el margen que puede, o no, sacrificarse.

### ***Capítulo 2 – Trabajos Relacionados***

Se divide en dos secciones, la primera menciona las fuentes más importantes para la ejecución de esta investigación, la segunda, dos ejemplos de problemáticas resueltas que tienen

---

puntos en común con nuestro problema. Se explican los puntos que hacen tan importantes dichas fuentes y los puntos de común y diferencias de los casos con el que nosotros proponemos.

### ***Capítulo 3 – Desarrollo***

Comienza con la presentación de puntos de vista, tanto el nuestro propio como otros posibles. Sin dejar de lado que los puntos de vista no son correctos o equívocos, sino distintos se justifica la postura de nuestro punto de vista. Luego, se presentan en forma detallada cada uno de los elementos que posee nuestra propuesta. Se brindan las justificaciones de cada uno de ellos y algunas consideraciones que no deben dejarse de lado en una posible implementación. Es aquí donde se presenta la arquitectura y la interacción de las partes, otorgando a cada una de ellas sus respectivas responsabilidades. Este capítulo explica con claridad la operatoria y razón de ser de cada elemento propuesto, siempre teniendo en cuenta lo estudiado dentro del marco teórico.

### ***Capítulo 4 – Discusión***

Nuestro último capítulo brinda un panorama macro sobre la propuesta en su completitud. Se presenta la ruta completa que involucra, brindando un detalle menor de cada sección o componente que el capítulo anterior. La lectura de esta sección ordena los conceptos y la lógica de la propuesta. Adicionalmente, se brindan algunas recomendaciones extras que, aun encontrándose fuera de los límites del flujo propuesto, son relevantes para una implementación práctica con resultados optimizados. En este capítulo son detalladas las conclusiones de nuestro trabajo, así como presentadas las líneas de investigación que se abren a partir del mismo.

## Capítulo 2 – Trabajos Relacionados

### Introducción

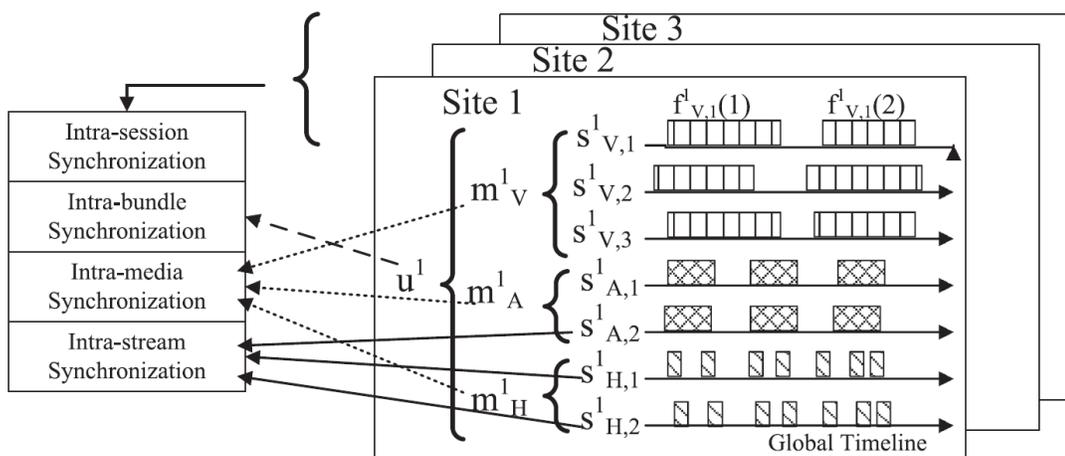
En este capítulo presentaremos, en primer lugar, los tres trabajos que creemos más importantes en el abordaje teórico de nuestra investigación. Dicha investigación es el fundamento que nos permite comprender el contexto completo de nuestra problemática para poder resolverla. Luego mencionaremos dos casos de sistemas ya implementados que creemos similares a nuestro proyecto. No obstante, se tiene en cuenta que los enfoques y necesidades de ellos no se aplican completamente a nuestro caso.

### Fuentes Relevantes

“Evolution of Temporal Multimedia Synchronization Principles: A Historical Viewpoint” (Huang & Nahrstedt, 2013) es un trabajo con gran utilidad para la comprensión del modelo multimedia y sus tipos de sincronización. Para nuestra investigación resulta relevante, no solo por su recopilación histórica, sino, por su claridad al presentar, tanto el modelo de jerarquía multimedia, como el modelo de cuatro capas de sincronización se observa a continuación.

### Figura 5.

*Cuatro capas de sincronización presentes en Huang & Nahrstedt.*



*Nota.* Tomado de *Four layers of synchronization relations* (pág. 3) por Huang & Nahrstedt, 2013, *Evolution of Temporal Multimedia Synchronization Principles: A Historical Viewpoint*.

---

Si bien en nuestro trabajo presentamos las capas de sincronización de una forma diferente, nos nutrimos en gran medida de esta fuente. La mayor diferencia con ella es el tener en cuenta diferentes formas de nombrar cada una de las capas. Existen varios autores que mencionan de uno u otra forma la sincronización de alguno de estos niveles, pero no existe una convención definitiva sobre la manera correcta de referirse a ellos. Para no confundir conceptos al leer distintas bibliografías, debemos tener en cuenta cuando se habla de intra (dentro) o inter (entre), y en qué nivel jerárquico se sitúa dicha palabra. Una vez situados allí, tener en cuenta que, en el caso de intra, puede referirse o incluir los niveles jerárquicos inferiores y, en el caso de inter, sucede lo mismo con los niveles superiores.

Otro trabajo que deseamos destacar por ser rico en su contenido, para nuestros propósitos, es “Impacto de Parámetros de QoS en Aspectos de QoE: Análisis desde el Punto de Vista de la Sincronización Multimedia” (Montagud, Boronat, Martínez, Belda, & Cesar, 2015). En este trabajo también se mencionan distintos tipos de sincronización y, a diferencia del trabajo del párrafo precedente, menciona también las sincronizaciones entre varios remitentes o destinatarios. El recorrido del escrito muestra cómo y qué pueden afectar a la sincronización, esto resulta muy útil para nuestros propósitos. Adicionalmente, también encontramos allí varios datos cuantitativos, específicamente en milisegundos, sobre los factores que menciona.

Durante el relevamiento de fuentes realizado para el desarrollo de nuestro trabajo, en todo momento se tuvo en cuenta la necesidad de establecer una referencia fiable a partir de la cuál realizar la sincronización. Si bien no se ha citado aún en nuestro trabajo, Montagud, Boronat, Martínez, Belda, & Cesar, mencionan que “la primera decisión consiste en seleccionar la referencia temporal a la que sincronizarse” (2015, pág. 7). Esto no solo confirma la necesidad que se tiene en mente, sino, que nos orienta en el camino que debe seguirse tanto en nuestro trabajo como en cualquiera que fuese basado en la sincronización de archivos multimedia.

El tercer trabajo que deseamos destacar en este apartado es el realizado por Haining Liu y Magda El Zarki. Para esto referenciamos, en principio, a Liu & Zarki (2003), de todas formas, deseamos que se entienda a *su trabajo* como la consecución de escritos que han publicado y pueden encontrarse en nuestra bibliografía. En ellos presentan un esquema de control para la transmisión en vivo de contenido multimedia, basado en la demanda de contenido hacia un servidor. Su trabajo es de gran interés ya que presentan un mecanismo de almacenamiento en buffer que, de ser necesario, se adapta de manera dinámica para mejorar la calidad de

---

experiencia del consumidor. Volveremos a referirnos a esto en el apartado que comienza en la página 58, El Buffer.

### **Casos Similares**

Durante la realización de nuestro trabajo hemos encontrado dos casos que consideramos similares al que hemos planteado para nuestra investigación. Éstos son el caso de Ságora y el de la transmisión de deportes que realiza Disney.

“SÁGORA es una plataforma que permite conectar usuarios mediante audio transmitido en alta calidad y muy baja latencia. Es ideal para la práctica musical y educativa ya que en SÁGORA no existe cancelación de señal. Pueden emitirse y oírse dos o más señales de audio al mismo tiempo.” (sagora, 2021)

Se trata de un software libre que puede ser descargado directamente de su sitio oficial. El mismo fue desarrollado por un grupo de investigadores de la Universidad Nacional de Quilmes. Permite a los músicos realizar ensayos de forma remota, reduciendo en la mayor medida posible las latencias entre ellos, impactando de igual manera en la percepción del sonido. Su principal diferencia con nuestro proyecto es la decisión de utilizar únicamente audio para poder lograr sus objetivos de ensayo en tiempo real. Esta información no solo puede consultarse en sus medios oficiales, sino, que fue mencionada en el programa de Vorterix (2022) del 25 de junio de 2020. En el mismo, se presenta una nota realizada a Diego Romero Mascaró, docente, músico, investigador y director de la Escuela Universitaria de Arte de la Universidad Nacional de Quilmes.

El caso Ságora, si bien se diferencia del nuestro, enfocándose en audio y no video, para poder brindar la posibilidad de ensayar en tiempo real oyendo a cada uno de los compañeros participantes, se asemeja en su necesidad musical. Por otro lado, el caso Disney, no tiene involucrada ninguna característica musical, pero sí de sincronización y diferencias de retardos.

Este caso se comenta en la entrevista realizada, también por Vorterix (2022) a Fernando Tornello, el 6 de octubre de 2021. Fernando es relator de Fórmula 1 y comenta que antes de la pandemia de COVID-19, la transmisión la realizaba, mayormente, desde el canal, en Argentina. Esta transmisión era sumada a la transmisión de las propias carreras y comentarios de una persona que siempre se encontraba en el circuito. Así como se sumaba una transmisión desde México de igual forma que la de Argentina. El resultado era una correcta sincronización entre la

---

carrera y quien se encontraba en la pista, con 1 o 2 segundos de retraso para los comentarios desde Argentina y 3 o 4 para los que se agregaban en México.

Disney realizó una inversión de aproximadamente 25 millones de dólares para lograr realizar transmisiones completamente remotas, con cada persona en su domicilio particular, y resolver adicionalmente el problema de sincronización entre ellos. Actualmente, reciben una transmisión preferencial de la carrera 15 segundos antes de que la misma llegue a los consumidores finales de televisión. Si bien el retraso desde la adhesión de los relatores y comentaristas hasta la visualización en televisores es de 15 segundos, existe otra transmisión a usuarios finales, que llega 3 o 4 segundos antes, por medio de una aplicación que se conecta a internet. Esta aplicación agrega una serie de estadísticas que son consultadas también por los relatores, con lo cual un usuario que visualiza, tanto la aplicación como la televisión, podría oír dichos datos 3 o 4 segundos después de haberlos visto en la aplicación.

El caso Disney ha logrado sincronizar transmisiones de distintas partes del mundo, sin que los proveedores deban salir de sus hogares, adicionando entre 11 y 15 segundos, sea por aplicación o televisión, hasta la visualización final por los usuarios consumidores.

---

## Capítulo 3 - Desarrollo

### Introducción

Habiendo comprendido el problema a resolver y el marco teórico que lo rodea, se da lugar al desarrollo propio de la temática planteada. En él, pretendemos clarificar el punto de vista a adoptar para su ejecución y presentar los distintos componentes que serán necesarios para la solución del problema, así como la interacción que pueda existir entre ellos.

### Puntos de Vista

Resultaría inviable la transmisión de un concierto completamente en vivo, sin ningún milisegundo de retraso, o con tan pocos que se pueda interpretar que no habrá repercusiones en la calidad del mismo. La única forma de considerarlo viable y, al menos analizarlo, sería poder garantizar que todos los paquetes lleguen a destino en un tiempo mucho menor al rango de percepción humana, por ejemplo, de 3 milisegundos. Siempre que existe un tiempo de demora existirán, a la larga, desfases generados por el efecto de bucle que se puede generar entre las distintas partes. Este efecto de bucle se refiere a cuando el músico  $\alpha$  recibe el sonido de  $\beta$ , por ejemplo, 3 ms tarde. Suponiendo que la percepción, reacción y anticipación humana fueran perfectas,  $\beta$  responde a  $\alpha$  en el mismo instante, pero, aun así, llega con 3 ms más de demora. En esta instancia  $\beta$  recibió con 6 ms de demora. Al responder, llegará a  $\alpha$  con 9 ms de demora, y así sucesivamente. Esto producirá que en cierto punto comience a ser perceptible. Los músicos intentarían ralentizar su interpretación para acoplarse a lo que oyen, pero esto sucederá posiblemente en ambos extremos. Extendiendo esto a una orquesta de una gran cantidad de integrantes podría haber un gran caos. La figura del director o concertino evitaría este caos, ya que todos seguirían a esa persona como guía sobre la cual interpretar su parte de la obra. No obstante, todos tocarían con 3 ms de demora y llegaría a los consumidores finales de la obra con 6 ms de demora.

Dada la situación anterior, podemos imaginar que no es sencillo seguir solamente a quien dirige cuando se oyen otros sonidos que no van a tiempo. Para evitar esto, debería quitarse a los músicos la posibilidad de oír al resto de sus compañeros. En cuanto a los consumidores finales, podrían notar el desfase entre quien dirige y el resto de los músicos. Una vez más, para evitarlo, podría quitarse su visualización por parte del público final, y que solo llegue a ellos lo que tocan el resto de los músicos. Entiéndase músicos como una generalización del ámbito de

---

este trabajo, que podría bien ser reemplazados por bailarines o actores para otras finalidades. Otra opción, sería calcular ese desfase exacto para retrasar la visualización de quien dirige y así acoplarlo al resto de intérpretes en la visualización final. Estas dos opciones mencionadas no son nuestra recomendación, sino, que las mencionamos para idealizar las posibilidades que nos ayudan a concluir con nuestra propuesta.

En primer lugar, para ambas posibilidades mencionadas en el párrafo anterior, todo parte de un supuesto no real. En el mismo no existen tirones en la red y se mantiene constantemente a una demora demasiado baja, de 3 ms, lo cual no es normal en las redes actuales. Desde cualquier computadora puede comprobarse que al enviar una prueba de *ping* a cualquier sitio no se recibirá una respuesta tan rápida. Que tan rápida sea dependerá de muchos factores y no en todos los equipos y redes será igual, pero sería un caso muy extraño encontrar los 3 ms. El lector puede comprobar esto en este preciso momento realizando una prueba en su propio equipo. No es nuestra intención brindar un tiempo estimado, ya que los fines de este trabajo son inclusivos en cuanto a las disponibilidades de cada persona.

En segundo lugar, al tomar la opción de escuchar al resto de intérpretes a quienes deben ejecutar las obras, no estamos brindando un mayor apoyo que el que proponemos en este trabajo. De igual manera, al diferir en el tiempo al director, para la visualización final, tampoco estamos evitando el diferir en el tiempo para lograr el objetivo. De hecho, no pudiendo suponer realmente que todos los ejecutantes tendrán la misma demora, habría que comenzar a diferir uno por uno de maneras diferentes, lo que termina generando más trabajo que el planteado.

Debemos dar crédito igualmente a los creadores de SÁGORA, quienes, partiendo de un punto de vista similar, crearon una aplicación (sagora, 2021) que permite ensayar a los músicos de forma remota. Las diferencias de esta aplicación con nuestra propuesta son, principalmente, que en todo momento involucra solo audio, en ningún momento entra en escena el video y, por momentos, puede notarse la pérdida de fragmentos del audio. El objetivo no es una gran calidad en el audio, sino en el tiempo de ejecución. Fue realizada buscando la manera más rápida de entregar los paquetes entre los distintos participantes, omitiendo la reproducción de aquellos que no llegan en el tiempo deseado. De esa forma, los músicos realizan sus ensayos remotamente sin dejar de oír el ritmo completo de sus compañeros, apoyados en sus capacidades de predicción del tiempo, sin necesitar oír todas las notas ejecutadas para poder continuar con la interpretación. De extender esto a un público general en forma de concierto, y sin agregar mayor retardo en la

---

reproducción, se notaría una calidad de sonido disminuida, con pequeños cortes en la interpretación de cada persona.

Otro posible enfoque sería el de analizar en tiempo real cada flujo de audio y video e ir corrigiendo sus tiempos de reproducción, de forma artificial, para lograr la perfecta sincronización. Esta idea surge naturalmente a raíz del trabajo de Román Sarmiento (2020), titulado “Detección de Sincronización Audiovisual Utilizando Deep Learning”. En principio, esta herramienta conllevaría un procesamiento de datos que suma demoras por tiempo de proceso, con lo cual también estamos hablando de diferir en el tiempo el envío al usuario final. También creemos que quitaría la calidad de experiencia realista que estamos buscando, lo cual no es completamente una contra o un error, sino, un enfoque distinto. No obstante, debemos mencionar para este caso la posible existencia de desincronizaciones voluntarias en una obra. Como lo indica Kopiez, “es claro que la capacidad para des- sincronizar es igualmente importante por la variación entre exacta sincronización e intencional asincronía producida en el 'toque humano' de la interpretación musical" (2002, pág. 534). Lo que en realidad simularía esta herramienta es la edición manual que realizan hoy en día los editores de videos. Como se mencionó al plantear nuestro problema, se editan los videos manualmente para unificar, tanto imágenes, como sonidos, corrigiendo las diferencias temporales de cada audio e imagen.

### **El Buffer**

Un actor principal en nuestra arquitectura propuesta es el *buffer*. Se trata de una palabra que utilizaremos en inglés, la cual podríamos traducir como amortiguador o llamarla colchón, aunque también podemos encontrarla en traducciones como búfer. Generalmente representa un determinado espacio que se reserva para alojar archivos aún no transmitidos o reproducidos, aunque esto podría ya haber ocurrido. En nuestra propuesta, por ejemplo, se incluye un buffer donde se guardan paquetes que pueden ya haberse transmitido, por si fuesen necesarios nuevamente por algún cliente en particular. Al almacenar cierto tiempo de reproducción del contenido en memoria, podemos luego mostrarlo aun cuando no se estén recibiendo paquetes, por periodo de tiempo, debido a inestabilidades de la red. La existencia del buffer se fundamenta en la necesidad de lograr la sincronización multimedia “a pesar de la existencia de retrasos y diferencias de retraso a lo largo de la cadena de distribución de extremo a extremo” (Montagud, Cesar, Boronat, & Jansen, 2018, pág. 8). Estos autores lo mencionan específicamente para la

---

sincronización intra-media. En otras palabras, una función fundamental es evitar los problemas que pueden generar los tirones, conocidos como jitter.

Para lograr este tipo de sincronización en sistemas multimedia distribuidos, las estrategias de almacenamiento en buffer de reproducción se emplean normalmente en el lado del receptor para suavizar los efectos del jitter o la variabilidad del retardo (e incluso de la pérdida de paquetes, mediante la adopción de técnicas de recuperación o retransmisión de errores). Por un lado, el tamaño del buffer de reproducción debe ser lo suficientemente grande para compensar los efectos del jitter. Por otro lado, los retrasos de almacenamiento en buffer deben ser tan cortos como sea posible para minimizar la latencia del servicio multimedia. (Montagud, Cesar, Boronat, & Jansen, 2018, pág. 8)

Para nuestro objetivo no resulta perjudicial la existencia del buffer mientras sea de algunos segundos, o hasta unos pocos minutos si fuese necesario y aceptado por los usuarios. Esto no afectará la calidad de experiencia del usuario final que se desea lograr. Dentro de las posibles contras, podrían darse casos donde comience a ser más relevante este retraso planteado. Por ejemplo, si deseara agregarse una función de chat, produciendo que los mensajes se envíen reaccionando a situaciones que en la realidad ya ocurrieron poco antes a lo visto. Si se utilizara la aplicación para interactuar con el público podría notarse estos retrasos, más que nada, si se extendieran a minutos. Lo que realmente deseamos evitar es retrasar la reproducción durante horas para grabar el concierto completo antes de transmitirlo. Hacer esto quitaría la sensación de realismo, por ejemplo, al ver una transmisión que posee luz diurna, mientras el consumidor sabe que los intérpretes están tocando en su misma ciudad, y la transmisión se realiza durante la noche.

### ***Buffer Dinámico***

Adicionalmente al típico buffer mencionado, se utilizan técnicas para alargar este tiempo, por ejemplo, disminuyendo la velocidad de reproducción. Esto puede apreciarse, entre otros, en el trabajo titulado “Towards the Delay and Synchronization Control for Networked Real-Time Multi-Object Multimedia Applications” (Liu & Zarki, 2003, págs. 2-3). Esta técnica puede producir un efecto notorio al ojo y oído humanos si no se la limita correctamente. Para nuestro

---

caso, teniendo en cuenta la existencia de la falla humana que pudiesen incluir los contenidos audiovisuales, podría significar la diferencia entre perceptible y no perceptible. Por ejemplo, si un músico demoró 25 ms más que el resto en dejar de reproducir un sonido largo. Si se tomara ese sonido largo para completar un espacio, alargando su reproducción otros 25 ms, la suma de retraso sería mucho más apreciable. Esto puede ocurrir, aunque no hubiese error humano, ya que no conocemos la cantidad de tiempo de duración de los paquetes potencialmente perdidos. En este último caso, igualmente, sería más sencillo limitar el tiempo disponible a unos pocos milisegundos para alargar reproducciones, teniendo en cuenta que se conocería el acierto humano con error 0 ms.

Consideramos una mejor opción, en lo que respecta al audio, el omitir los sonidos no recibidos, al menos, cuando la transmisión posea más de un músico. No obstante, podría aceptarse la utilización de dicha técnica en el cliente final. De esta forma, si el mismo posee problemas de red, recibirá un contenido con mayor fluidez al que podría recibir en caso de omitir la reproducción de los cortos tiempos sin recepción. Una vez más, la cantidad de milisegundos para esta técnica debe ser limitada y, si se supera, recurrir nuevamente a la omisión. Pensando en la percepción y la calidad de experiencia del consumidor final, el no oír tocar un instrumento podría ocasionar 2 escenarios. Por un lado, cuando no haya más músicos tocando el mismo instrumento y la misma nota. En este caso podría creerse que el músico olvidó tocar dicha nota, o la tocó durante menos tiempo, cuando se trata de milisegundos, o suponer también que hay problemas de red cuando esto durase un periodo más largo, por ejemplo, segundos. Por otro lado, si existen más músicos ejecutando la misma interpretación, el omitir a uno de ellos no produciría errores notorios. Cuanto mucho, podría notarse un volumen ligeramente disminuido de dicho instrumento, al oír, por ejemplo, a 4 violinistas en lugar de 5.

En lo que respecta a las imágenes del video, es más factible el reproducir fotogramas durante más o menos tiempo, de acuerdo con las necesidades del buffer de reproducción del consumidor final. Oír un sonido completamente fluido, natural y realista, es más importante que la visualización perfecta del video. Esto lo mencionaremos en el apartado de Sincronización, de la página 70.

Sea el caso que fuere, sonido, imagen o ambos, desde el ámbito del servidor, al no recibir un paquete determinado o recibirlo con algún faltante, transmitiremos de la misma forma que fue recibido. Es decir, se utilizará la omisión. Esto no es con objetivos de negar las técnicas

---

mencionadas, sino, de dejar dicha responsabilidad al consumidor del servicio. Esto ayudará a evitar inconvenientes más notorios de lo deseado. En caso de rellenar espacios vacíos desde el servidor y entregar un flujo simuladamente continuo, quien lo reciba no sabrá que ya fue manipulado. Podría darse el caso de que al recibirlo se vuelva a utilizar la técnica en el mismo fragmento, excediendo los límites establecidos sin saberlo. En cambio, al retransmitir el error, y si existe un error adicional en la descarga del consumidor, el mismo sabrá la duración total del faltante, para poder decidir correctamente la estrategia a emplear.

### *¿Dónde Utilizamos el Buffer?*

A ciencia exacta, no pretendemos la utilización de un buffer, sino, de varios de ellos. En principio, sabemos que habrá varios puntos de envío de contenido en tiempo real hacia el servidor centralizador. Cada uno de estos puntos lo consideraremos proveedor, así como al punto final lo hemos considerado hasta aquí consumidor. En este fragmento, el servidor centralizador sería el cliente de la relación, no obstante, y a fines prácticos, mantendremos la nomenclatura proveedor, servidor, consumidor. Dicho esto, nos referimos entonces a la relación proveedor-servidor. Será necesario mantener un buffer en cada proveedor, que guarde los paquetes que el servidor podría aún requerirle. Fuera de creer que el envío será en tiempo real, y aunque inicie de esta forma, el envío podría tener que repetirse en algunos paquetes y no en otros, mezclando el orden real del mismo. Cada proveedor deberá almacenar en un buffer de salida los paquetes que el servidor aún necesite.

El servidor en sí, como intermediario, deberá tener un buffer diferente, al que podríamos llamar buffer compuesto. En primera instancia, recibirá flujos de varios proveedores, por lo cual almacenará algunos mientras aún no recibió otros. La transmisión al consumidor debería iniciar sabiendo que todos fueron recibidos. Este es el primer buffer que necesita el servidor, para alojar durante un tiempo los paquetes de algunos segundos (o milisegundos) de todos los proveedores, previo a poder entregarlos. Una vez que este buffer se encuentra lleno, o transcurrió el tiempo previsto, comenzará el envío de la relación servidor-consumidor. Para garantizarla, habrá una segunda parte de este buffer, que tendrá la misma finalidad que el buffer de salida que se mencionó en el proveedor. Es posible que algunos consumidores ya hayan recibido ciertos paquetes, pero otros no, sea por una demora en pedirlos o por solicitarlos nuevamente por alguna falla. Para garantizar esto, debe contemplarse un sector de buffer para los paquetes que teóricamente ya habrían transcurrido en el tiempo.

---

Finalmente, tendremos un buffer de reproducción en el consumidor final. Éste es un buffer que no necesariamente debe llenarse antes de comenzar la reproducción. Al existir los archivos ya alojados en el servidor, un consumidor, con una velocidad de descarga suficiente, podría descargar archivos más rápido de lo que los reproduce. De esta forma, con recibir unos pocos paquetes sería suficiente para iniciar la reproducción, e ir llenando el resto del buffer con el transcurso del tiempo. Lo que será un limitante de este buffer es la disponibilidad del servidor para enviar los archivos, ya que no se podrían descargar con una velocidad que supere la disponibilidad del servidor. Este buffer es en el cual se tendrá una mayor libertad para refaccionar errores, decidiendo cuando omitir reproducciones y cuando rellenarlas, alargando o adelantando una reproducción.

### ***¿Cuánto Tiempo Almacenar y Qué Impacto Genera?***

En principio debemos tener en cuenta que no existe una fórmula perfecta para determinar el tamaño de un buffer mientras el entorno de red no sea completamente predecible. Si bien se plantea en esta arquitectura la existencia de una tarea de prueba de red, eso no garantiza ni una red estable, ni tampoco que sus inestabilidades sean las mismas durante un largo periodo de tiempo. Un buffer, cuanto más largo sea, más aumentará la estabilidad de reproducción, disminuyendo la cantidad de cortes, adaptaciones (como la ralentización de reproducción para llenar vacíos) y cambios de resolución. Cuanto más corto sea, menor impacto tendrá en el retraso de extremo a extremo. Adicionalmente, un buffer corto consume una cantidad menor de transferencia de datos cuando un usuario decide no visualizar el contenido completo. De todos modos, esta última aclaración, en nuestro caso, solo vale para el buffer del consumidor final. En cuanto al impacto de extremo a extremo, veremos también que no todos los buffers propuestos afectarán.

Es preciso diferenciar dos conceptos temporales. Uno, es la cantidad de tiempo de reproducción que contiene un segmento determinado. Otro, la cantidad de tiempo que demora la transferencia de dicho segmento. Traemos, a modo de ejemplo, algunos datos del apartado experimental del trabajo de Mata Galiano (2014). En su trabajo se puede observar, en más de una oportunidad, demoras de entre 400 y 550 milisegundos en descargar conjuntos de segmentos de audio (96 Kbps) y video en 1280x720 (2.5 Mbps). Adicionalmente, podemos observar demoras de menos de 150ms cuando el video es en 320x180 (entre 300 y 350 Kbps), combinados con un audio idéntico al anterior, de 96 Kbps. Todos estos segmentos tienen una duración de 2 segundos

---

de reproducción, con lo cual podemos notar la diferencia entre los tiempos indicados. Si bien brindamos algunos ejemplos posibles, las pruebas de entorno serán las que determinen cuantos milisegundos se tarda, aproximadamente, en obtener un paquete o segmento completo de todos los participantes.

Introducimos, apoyándonos en el trabajo de Mata Galiano (2014), el concepto de tasa de bits (en inglés *bit-rate*), utilizado para mencionar las velocidades de descarga. De esta forma, no se mencionan en dicho trabajo, por ejemplo, 2.5 Mbps, sino 2500000, lo que se comprende en bits por segundo. Recomendamos, a fines prácticos, el intentar mantener estos dos tiempos mencionados lo más similares que sea posible, así como también procurar que el tiempo sea lo más corto que se pueda. En este acercamiento de valores, debe procurarse que el tiempo de descarga no supere nunca al de reproducción. Si esto ocurriese, el buffer siempre irá perdiendo tiempo de reproducción, por más que sea lentamente por una diferencia muy pequeña, pero de manera constante, produciendo un vaciamiento. En el mismo trabajo mencionado se indica, en más de una ocasión y en diferentes protocolos, que 2 segundos es el menor tiempo de duración que se puede establecer a un segmento. Utilizaremos entonces 2 segundos como duración de referencia.

Centrándonos en el servidor que almacenará y distribuirá el contenido, sin excesos incoherentes, creemos preferible un margen amplio y no uno demasiado ajustado. Indicaremos lo que consideramos el mínimo necesario, pero siempre se logrará aún mayor estabilidad adicionando tiempo. Esto es una decisión que preferimos dejar abierta a quien utilice o implemente el sistema, ya que será conocedor de las necesidades reales en su caso aislado. La relación de transferencia entre el proveedor y este servidor será, en principio, la que más asemeje el tiempo a los 2 segundos de reproducción mencionados. Tomando como ejemplo los datos los párrafos precedentes, podemos tomar 550 ms como tiempo de descarga del contenido en 1280x720 (conocida como 720p) y audio, dejándonos un margen de 1450 ms. Estos podrían bien completarse con la solicitud de más resoluciones, probablemente sea suficiente para solicitar las 4 resoluciones menores que daremos como ejemplo en el próximo párrafo, o ninguna menor, pero sí una de 1080p. Otra alternativa, será la de completar, en ese tiempo, las resoluciones faltantes por cuenta del servidor (o tercerizando). Sobre esto hablaremos más adelante, pero es importante comprender aquí la forma en que se utilizarán los tiempos para lograr ese semi

---

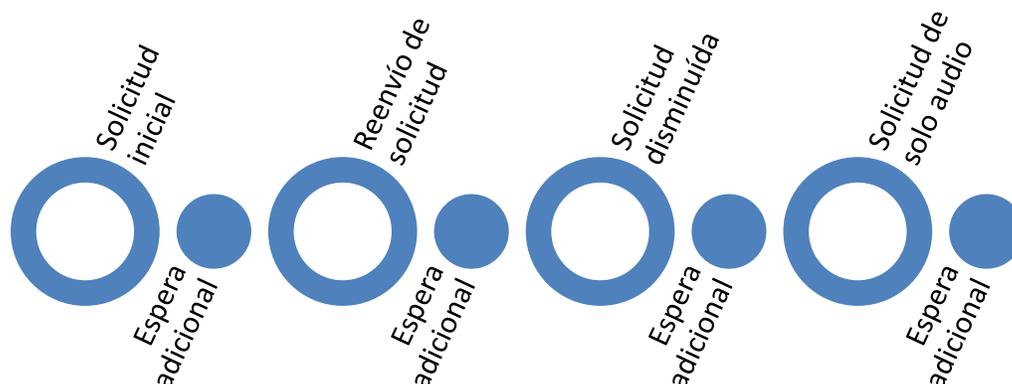
emparejamiento. Al dejar un poco más corto el tiempo de transferencia que el de reproducción, también se tienen en cuenta tiempos de procesamiento.

La reducción del tiempo de disponibilidad del buffer, por no estar recibiendo paquetes en tiempo y forma, debería ser minimizada en la mayor medida posible. No confundir esto con el uso de los archivos del buffer. En la reproducción normal se utilizan los archivos del buffer constantemente, lo que debe ocurrir es que se vuelve a llenar, con la misma frecuencia, con un nuevo segmento. Para evitar el vaciamiento del buffer, el servidor gestionará la calidad y cantidad de contenido que solicita a cada uno de los emisores. Ejemplificando, un músico  $\alpha$ , que posee una conexión estable y veloz, podría enviar contenido tanto en solo audio como en 144p, 240p, 360p, 480p, 720p y 1080p al mismo tiempo. Esto dependerá también de las capacidades de su hardware para transformar el video, no solo de su entorno de red. La transformación del video a la que referimos es la transcodificación, a la cual haremos referencia en la página 67. Por otro lado, un músico  $\beta$ , con una conexión de menor ancho de banda y/o más inestable, podría enviar contenido de solo audio y 720p únicamente. En aquellos momentos que deje de recibirse el contenido a tiempo comenzaría a disminuir el buffer, pero inmediatamente se comenzaría a solicitar el contenido en 480p, en lugar de 720p, gestionando estas rebajas hasta limitarlo a solo audio, en casos extremos. Tener en cuenta que la calidad del video no está dada únicamente por la cantidad de píxeles, sino también por la cantidad de cuadros por segundo. Es posible que un contenido se grabe en 720p a 60fps y se transmita al servidor, tanto a 60fps como a 30 fps. Un video de mayor calidad implica la necesidad de una mayor tasa de transferencia, o tasa de bits. Dicho esto, podemos adentrarnos un poco más en cuánto tiempo se necesitará, mínimamente, para un buffer efectivo en el servidor.

Recomendamos que el buffer del servidor posea hasta cinco segmentos de duración. Esto se fundamenta en la lógica de solicitudes que se realizarán, siguiendo el orden a continuación graficado. Otra alternativa, si se desea una mayor reducción del retraso de extremo a extremo, con un mayor riesgo, pueden ser hasta cuatro segmentos, eliminando la segunda solicitud graficada y correspondiente espera adicional (reubicando ese tiempo como se comprenderá en el próximo párrafo). Claro está que puede reducirse aún más, pero de esta forma evitamos riesgos innecesarios, ya que no solo deseamos un retraso mínimo, sino la garantía de la transmisión. Notar el uso de la palabra hasta, ya que, como veremos luego, hasta cinco podrían transformarse en cuatro de forma dinámica.

**Figura 6.**

*Esquema de solicitudes propuesto.*



En primer lugar, se realiza una solicitud inicial con la calidad, o calidades, determinadas previamente. La determinación es realizada inicialmente por la prueba de entorno de red, y puede ser aumentada o disminuída durante la transferencia, de acuerdo con las posibilidades (vaciamiento o desbordamiento de buffer). Esta solicitud tiene un tiempo definido en el cual esperamos recibirla, el cual ya comentamos. Al mismo, se adiciona un margen de error, en este caso del 25% si se sigue nuestra recomendación del quíntuplo. Luego de transcurrido el tiempo, si aún no se ha recibido correctamente, se repite la misma solicitud, y se aguarda la misma cantidad de tiempo. Persistiendo la no recepción, o recepción con fallas, se realiza una solicitud de una resolución menor, o menos cantidad de resoluciones, y se aplica la misma lógica de espera. Finalmente, se resigna la imagen y se solicita únicamente el audio para intentar garantizarlo. Si en esta instancia continúa sin llegar, se omitirá también el envío posterior al consumidor, es decir, se deja vacío este espacio. Esto afectará entonces al retraso de extremo a extremo en un quíntuplo de lo determinado por la prueba de entorno. Cuatro partes de espera convencional y una quinta repartida entre estas cuatro, en partes iguales.

No se esperará siempre a la recepción de un segmento para comenzar a pedir el siguiente, sino, que se aguardará la duración del segmento para hacerlo. Con nuestro ejemplo recomendado de 2 segundo (el cuál recomendamos reducir aún más si el protocolo utilizado lo permite), habrá que esperar, lógicamente, 2 segundos para solicitar el próximo. Esto se debe a que el proveedor

---

tardará ese tiempo entre uno y otro segmento en realizar el contenido. Con lo cual, en un caso pesimista, donde deba acudir a la segunda solicitud del primer segmento, se habrá realizado ya la primera solicitud del segundo segmento. Puesto en ejemplo práctico, en el segundo 0 se realiza la solicitud 1 del segmento 1, en el segundo 2 la solicitud 1 del segmento 2, y en el segundo 2,25 la solicitud 2 del segmento 1. De esta forma, si el caso no llegase hasta la última instancia, en la cual se solicita solo audio, tendremos el primer segmento completo antes de transcurridos los primeros 10 segundos. Si esto sucede también con los próximos 2 segmentos (es decir con los primeros 3 segmentos), se habilitará la posibilidad de transmisión acortando el buffer a 4 partes en lugar de 5. Es decir, tendremos un retraso de 8 segundos en nuestro ejemplo y no de 10. No recomendamos acortarlo aún más, ya que inevitablemente habrá errores con el correr del tiempo de reproducción y es necesario ese margen mínimo de amortiguación de fallas.

En cuanto al buffer negativo, o de archivos probablemente ya transferidos, no afectará en absoluto la demora de transferencia, sino, que representa un resguardo para los casos en que algún consumidor se haya retrasado o solicite reenvío. Su duración podría tener como mínimo la misma lógica que la ya mencionada, y no ser menor al buffer que posea luego el consumidor.

El consumidor podrá adoptar también la misma lógica que el servidor en su buffer inicial. En su caso sí influirá al retraso en la reproducción, con lo cual sumaría hasta un quintuplo extra. No obstante, la descarga del consumidor puede ser más veloz que el tiempo real de filmación, con lo cual su buffer se llenaría en menos tiempo. Esto puede producirse en esta instancia, pero no en la del servidor. El servidor recibirá los primeros paquetes, y todos los siguientes, con la limitación del tiempo que tarda en grabarse, es decir, no puede recibir en un tiempo determinado un mayor tiempo de reproducción. En cambio, cuando ya posee sus unidades de tiempo completas y se dispone a distribuirla, el consumidor podría descargar todos los paquetes en un tiempo mucho menor. Luego de esa descarga inicial, comenzaría la limitación de no poder descargar algo que todavía no posea el servidor. Esto depende, una vez más, del tiempo de creación del contenido del proveedor. Esta posibilidad otorga un beneficio extra al consumidor que, en el caso de experimentar demoras momentáneas, podrá nuevamente completar el buffer en un menor tiempo. Cabe destacar que el consumidor no realizará la descarga de múltiples calidades de video, pero sí de segmentos de múltiples usuarios en algunos casos. Aun así, su descarga adaptativa no debe sobrepasar su disponibilidad de visualización vía hardware. Con lo cual es mucho más factible la descarga en tiempos más acotados. Siguiendo los ejemplos

---

anteriormente mencionados, podría descargar los 2 segundos de reproducción en 550 ms. Cuando posea un buffer de la misma duración que el del servidor, comenzara a reproducir. En el mejor caso, podría implicar tan solo unos milisegundos, en el peor, entre 8 y 10 segundos, al igual que el servidor. Hasta aquí, entonces, tendremos 8-10 segundos de retraso en servidor, más 0-10 segundos en el cliente. Nunca será 0 segundos exactos, pero es posible que sea menos de 1 segundo. Entiéndase el tiempo mínimo a partir de aquí como no inclusive. Entonces, la demora será de 8 a 20 segundos.

Con respecto al proveedor, no posee en sí un buffer convencional que represente un tiempo de espera para el envío, pero deberá completar un segmento de grabación para poder enviarlo. Con lo cual, influirá con la duración de un segmento, en nuestro caso 2 segundos. Luego, guardará igualmente los segmentos, pero la lógica de este buffer es en negativo, al igual que el resguardo del servidor. El envío (por solicitud del servidor) debería realizarse inmediatamente al tener el paquete disponible, pero resguardarse en un buffer de salida para eventuales nuevas solicitudes. Este buffer también debe concordar con la cantidad de tiempo de espera del servidor, pudiendo garantizar el envío de un paquete desde de la solicitud inicial hasta la de solo audio. Transcurrido ese tiempo, por más que se recibiese una solicitud atrasada, el archivo podría haberse eliminado, ya que el servidor no lo estará esperando, habrá determinado omitirlo.

Finalmente, podemos indicar que la demora de extremo a extremo será de entre nueve y veintiún veces el tiempo de duración de un segmento. Con nuestro ejemplo, entre 18 segundos y 42 segundos. El peor caso difícilmente se dé en todos los consumidores, sus probabilidades de ocurrir en alguno de ellos son bajas, aunque puede darse en casos particulares.

### **Transcodificación**

Como anticipamos en la página 64, la transcodificación de videos se refiere a la creación de nuevos videos, a partir de los existentes, con características distintas, generalmente utilizado para la calidad del mismo en base a su resolución. Esto representa un proceso que, como tal, tendrá una cantidad de tiempo de ejecución, independientemente del responsable de realizarlo. van Deventer, Stokking, Niamut, Walraven, & Klos (2008) mencionan la contribución de la transcodificación de los flujos de video en el retraso de los mismo, e indican que esto afectaría entre 0 y 2000 ms. Está claro que en los ejemplos utilizados en la sección de buffers no podríamos ejecutar esta tarea si llegara a tardar 2 segundos por cada fragmento de video. No

---

obstante, estos valores no están indicados en la bibliografía para fragmentos cortos, de hecho, no se brinda información sobre a qué tipo ni duración de video se refiere, ni tampoco del hardware que lo ejecuta. Lo que nos permite saber, es que el valor de 2000 ms está indicado como un máximo posible en el peor escenario que hayan imaginado. En caso de que el hardware del proveedor demore realmente un tiempo inadmisibles, es preferible que envíe solo el flujo en calidad original, delegando la transcodificación a otra etapa. No es válido, por ejemplo, hacer fragmentos de mayor duración para evitar esto, ya que, lógicamente, no solo incrementa el tiempo de reproducción del fragmento, sino, el tiempo que demora la ejecución del proceso de transcodificación.

La tarea de transcodificación debe ser ejecutada por la primera parte posible en nuestro circuito de proveedor, servidor, consumidor. El consumidor es una última instancia que debería evitarse siempre que sea posible, ya que su descarga se vería dificultada si solo tuviese una opción de gran flujo de bits por cada proveedor. No obstante, debe tener la capacidad de hacerlo, ya que existen una gran cantidad de dispositivos y configuraciones de video diferentes. Aun si esto no fuera así, un usuario podría estar utilizando solo una parte de su pantalla para visualizar la aplicación. En ese caso, por más que se reciban resoluciones pequeñas, podría tener que adaptarlas, reduciéndolas aún más. Sería inviable tener una cantidad de resoluciones en servidor que satisfagan a todos los consumidores y en todo momento. Se debe tener, igualmente, una cantidad tendiente a satisfacer distintas capacidades de flujo de bits. Como indica Mata Galiano (2014, pág. 22), se recomienda que la diferencia del bit-rate, entre una y otra, sea de, al menos, un 50%, es decir, una relación de 1,5. De acuerdo con su recomendación, entendemos que, si fuese menor, no habría diferencias significativas en los tiempos de transferencias de una y otra. establecer una cantidad estándar de resoluciones posibles, que intentemos satisfacer antes del envío al cliente final. Creemos una buena práctica mantener siempre una relación de aspecto, por ejemplo, de 16:9, lo que incluiría 1920x1080 (1080p o Full HD), 1280x720 (720p o HD) y 640x360 (360p o nHD), entre otras.

La transformación a mayor cantidad de fps no será realizada, el cliente reproduce el doble de tiempo cada cuadro, o dos veces el mismo. En cuanto a la transformación a menor cantidad de FPS, sí debe ser tenida en cuenta. Para esto, se recomienda limitar también a un estándar, por ejemplo, 60 es múltiplo de 30, lo que permite una fácil transformación, tomando uno de cada dos cuadros. Estos valores suelen ser los que permiten las cámaras convencionales, aunque esta

---

condición podría cambiar a futuro. Lo aconsejable, es utilizar valores con algún múltiplo en común, por ejemplo, sería lógicamente más fácil la transformación entre 30, 60 y 75, que entre 30, 60 y 71. El consumidor podría también en este caso, reproducir la mitad de los cuadros sin transformación previa, pero no estaría aprovechando su capacidad de descarga de la mejor manera.

El servidor deberá tener también la capacidad, propia o tercerizada, de transformar resoluciones. Habrá casos en que la capacidad de red proveedor-servidor permita el envío de una calidad alta, pero no de varias en simultaneo. Asimismo, puede haber casos donde las capacidades de hardware local del emisor no le permitan realizar a tiempo la transcodificación efectivamente. En estos casos, al estar limitados al envío de una, o unas pocas, se priorizará la mayor calidad posible. Luego, se realizará la transformación, con el servidor como responsable, para tener una mayor cantidad disponibles para el envío. Si bien podría resultar más costoso hacer este trabajo en el servidor, y no delegarlo al receptor final, sería ilógico transmitir una calidad alta de cada emisor a todos los receptores. Cuando un receptor tenga una pantalla de 1920x1080, pero reciba 9 videos simultáneamente, los mismos podrían ser de 640x360 y llenar perfectamente la pantalla. El servidor central puede tercerizar la tarea de transcodificación, pero debe tenerse en cuenta el retraso de envío y recepción hacia y desde el servidor o servicio tercerizado. El flujo completo de la obtención del contenido no puede superar la duración de su propia reproducción.

### **Sincronización y Marcas Temporales**

Recordando lo estudiado en el marco teórico, sabemos que existen varios tipos de sincronización. Las de más bajo nivel son responsabilidad pura de quien genera el contenido en primera instancia, el proveedor. El mismo, debe garantizar que el contenido generado tenga buena sincronización dentro de los flujos (intra stream) y entre medios (inter media). Esto es, una correcta correlatividad de avance entre el tiempo real y la reproducción prevista para el contenido y una concordancia temporal entre su audio y su video. Para lograrlo, es importante que su reloj funciona correctamente, aunque no será relevante que se encuentre sincronizado con los del resto de proveedores.

Las sincronizaciones de mayor nivel derivarán de las marcas temporales colocadas por el proveedor en el contenido. Hemos dicho que no es relevante la sincronización de relojes entre

---

proveedores, esto es porque, como veremos, las marcas temporales no serán colocadas en base al tiempo actual de cada equipo.

### ***Guía Base***

Para comprender la manera en que colocaremos marcas temporales, es necesario tener en cuenta la pista base sobre la cual tocarán los músicos. Cada interpretación será realizada sobre una pista que podría ser tanto un audio, como un video, o un metrónomo programado con anterioridad. En caso de ser un audio o video, contiene ya sus propias marcas temporales, en base a las cuales es reproducido. Si se utiliza un metrónomo, anteriormente debe elegirse la velocidad y duración del mismo para la pista a interpretar, quedando como un archivo de datos en el servidor. Los proveedores deben realizar la descarga de estos archivos desde el servidor para, luego, poder tocar guiándose por su reproducción. Recomendamos que la descarga de todos los archivos incluidos en la transmisión se realice antes de la misma. Podría parecer una recomendación innecesaria, ya que anteriormente habrán realizado ensayos con las mismas pistas, pero de no ser así, se desea evitar la espera para poder iniciar la transmisión.

### ***Marcas Temporales***

Al realizar la interpretación en vivo, las marcas temporales que se colocarán en los archivos de audio y video generados, no serán las del reloj actual del equipo. Las mismas, serán copiadas del fragmento de pista base que se esté ejecutando. De esta manera, quedan automáticamente sincronizadas la pista base y el contenido del proveedor. La velocidad, o el sesgo, del reloj local de cada proveedor, no ocasionará desfasajes entre las mismas. En caso de haber algún milisegundo de sesgo en el reloj, la reproducción y la grabación incluyen el mismo sesgo, por lo cual no será notado en la ejecución del músico. Cada proveedor puede incluso tener sesgos diferentes, pero el copiar la marca temporal de la guía base asegura evitar errores futuros de sincronización. Adicionalmente, cada proveedor está acostumbrado al sesgo de su propio reloj, de todas formas, como se ha estudiado, la posibilidad de estos sesgos está alejada del rango de percepción, por lo que no afectará su capacidad de anticipación musical.

### ***Sincronización***

Como ya comentamos, las sincronizaciones dentro de los flujos y entre medios son responsabilidad del proveedor del contenido. Al mencionarlo no hemos incluido la sincronización dentro de los medios, esto es por considerar que cada medio tendrá un único flujo.

---

Si hubiese algún medio con más de un flujo, también es garantizado localmente por quien realiza la grabación. Tener también en cuenta que las nomenclaturas entre medio e intra paquete (intra bundle) refieren a la misma sincronización.

La sincronización entre paquetes, o dentro de la sesión, será la que el servidor debe garantizar, en la cual todos los proveedores estén coordinados entre ellos. Al haber utilizado las marcas temporales mencionadas, no hay un trabajo extra para el servidor, más que haber provisto a los proveedores, valga la redundancia, de las pistas bases, y entregar en conjunto todas las pistas a los consumidores. Ese momento, en el cual un consumidor solicita un segmento, entra en juego el momento de sincronización de este tipo, y se logra enviando el segmento de una misma marca temporal, pero de todos los proveedores.

La sincronización entre sesiones puede entenderse como la de garantizar que todos los consumidores del contenido lo reproduzcan al mismo tiempo. En este caso no se buscará dicha sincronización por la naturaleza de nuestros objetivos. Al no tratarse de una competencia, donde se brinda ventajas a quien lo reproduce algunos milisegundos antes que otro, carece de real importancia. De hecho, brindamos a cada consumidor la libertad de comenzar la reproducción en el momento que su buffer local lo permita. Esto de ninguna forma será exactamente igual para todos, ya que para esto deberían tener el mismo ambiente de red, mismo hardware y misma disponibilidad de recursos de ellos en un momento determinado.

Adicionalmente deseamos comentar que la sincronización de audiovisuales suele tener un maestro y un esclavo. Esto puede apreciarse en Liu & Zarki (2003) y sus posteriores trabajos. En nuestro caso, el maestro será siempre el audio, ya que deseamos la mayor exactitud posible en el mismo. Si debiera resignarse calidad de algún aspecto, siempre será la imagen del video la relegada, siendo en realidad un acompañamiento visual de la obra interpretada, no poco importante, pero sí menos importante que el audio. De hecho, si se hila fino, la sincronización del audio y el video no será exacta mientras existan tasas de refresco diferentes entre ellos. Probablemente el movimiento del arco de un violín filmado a 60 FPS se encuentre más cerca a la exactitud que la de uno filmado a 30 FPS, cuando se los contrasta con el mismo audio. Este fenómeno es producto de las capacidades técnicas del hardware y software, pero no corresponde a un error que se pueda percibir puntualmente, sino a la fluidez de cada video, a las cuales estamos acostumbrados. Tengamos también en cuenta que, en un concierto presencial, solemos escuchar el sonido más tarde del movimiento que percibimos gracias a la velocidad de la luz. “La

---

definición de sincronización correcta no es que un sonido correspondiente a una imagen se presente al mismo tiempo que la imagen. El contenido de las imágenes puede implicar diferentes retrasos del sonido respecto a la visión” (Salmon & Mason, 2008, pág. 8). Cuando la distancia entre el público y los músicos es considerable, también lo es esta diferencia, la cual es reducida en gran medida con los medios digitales propuestos. Salmon & Mason (2008, pág. 11) mencionan que, al ver una imagen bidimensional en una pantalla de una situación de la cual tenemos experiencias previas en el mundo real, dichas experiencias crean una expectativa de percepción. Si bien la pantalla se encuentra a una distancia diferente, la posible percepción esperada de dicha distancia queda anulada por la que esperaríamos en el ámbito real.

### **Pruebas de Entorno**

Así como mencionamos que las pistas guía deben descargarse a los equipos de los proveedores con anterioridad a la transmisión en vivo, diremos que la sesión interna debe comenzar antes que la transmisión pública. Esto, para tener un tiempo prudencial en el cual el servidor sepa que archivos puede brindar cada proveedor y cuáles no. En función de ellos, el servidor podría tener que generar, o tercerizar, las transcodificaciones correspondientes. Estos cálculos deben ser realizados en tiempo real en una sesión activa donde todos los proveedores participan. Durante la misma, se realizan las denominadas pruebas de entorno.

Las pruebas de entorno son en realidad una transmisión real, pero entre los proveedores y el servidor. Los proveedores inician la sesión y comienza el intercambio de paquetes con el servidor. Durante un tiempo prudencial, el servidor solicita cada vez más variantes de resolución y/o resoluciones mayores, hasta determinar cuál es el conjunto ideal para comenzar a llenar el buffer para transmitir públicamente. Una vez determinado, se envía la señal de inicio de la primera pista a todos los proveedores para que el contenido comience a ser generado.

Se debe tener en cuenta que estas pruebas no se realizan con uno, sino, con todos los proveedores. El fin es comenzar a llenar el buffer con segmentos útiles. En caso de no existir, debería comenzar solicitando una pista de resolución pequeña, con una tasa de transferencia de bits baja, luego, ir incrementando mientras las respuestas sean favorables y lo permitan. No deseamos que la transmisión comience con archivos de baja resolución únicamente, sino, de la mejor manera posible, ya que representarían la primera impresión dentro de la experiencia del usuario consumidor. De esta forma, las pruebas de entorno contribuyen directamente a la calidad de experiencia con el sistema.

---

Una consideración que debemos tener es la de la pluralidad de resoluciones que deseamos. A fines prácticos para el resto de las partes, es preferible que el emisor primario envíe varias resoluciones pequeñas, en lugar de solo una grande. Esto, teniendo en cuenta que no se ha pensado un concierto como la visualización en máxima calidad de un músico en forma individual. No obstante, podría brindarse la oportunidad de elegir hacerlo de esa forma al consumidor del contenido, aunque limitado a la mejor calidad que se posea de quien haya elegido. En la decisión, entra en juego la capacidad de realizar la transcodificación con el servidor como responsable, en caso de solicitar solo una resolución de mayor calidad. Si se dispone de los tiempos necesarios para hacerlo, será de esa forma.

Estas mismas pruebas pueden ser realizadas por el consumidor final para saber que resolución solicitar al servidor, no obstante, es libre se seleccionar manualmente la que desea. Esta opción es una decisión que se toma a la hora de generar las vistas del usuario consumidor. Así como será también una característica que podría o no estar en dichas vistas, la elección de ver a un músico en particular, a un conjunto de ellos, a todos ellos, a quienes estén reproduciendo sonido actualmente, o lo que se deseará brindar como opción de visualización.

## **Resumen**

Un actor principal en nuestra arquitectura propuesta es el buffer, que estará presente, tanto en el servidor, como en el consumidor y los proveedores, aunque con características diferentes. Otra tarea que deben poder realizar las tres partes mencionadas es la de transcodificación, cuanto antes sea realizada en la cadena, menor tiempo de retraso de extremo a extremo habrá.

El servidor gestiona las solicitudes que se realizan a los proveedores, partiendo de lo que determinen las pruebas de entorno, y cambiando de forma dinámica. Con el planteo que realizamos, la demora de extremo a extremo será de entre nueve y veintiún veces el tiempo de duración de un segmento.

La sincronización del contenido de distintos proveedores entre sí será realizada en base a marcas temporales, las cuales son copiadas de las pistas base hacia el contenido producido por cada proveedor.

---

## Capítulo 4 - Discusión

### Introducción

En el capítulo anterior, hemos presentado los componentes necesarios y propuestos para la resolución del problema planteado. En este capítulo, y en base a ellos, presentaremos un flujo que facilite la comprensión de sus interacciones y propósitos, apreciándolos de una manera ordenada y conjunta. Adicionalmente, mencionaremos algunas recomendaciones que, si bien se encontraban fuera de los límites de esta investigación, resultan de interés, o a tener en cuenta, para la ejecución de un proyecto basado en la misma. Finalmente, brindaremos nuestras conclusiones y las líneas de investigación que se abren a partir de la ya realizada.

### Flujo Propuesto

Los archivos de pistas bases son subidos, o configurados en caso de metrónomo, por el director de la orquesta, o quien cumpla ese rol en grupos musicales de otra índole. El mismo, es descargado por todos los integrantes y utilizado para los ensayos. Estos ensayos pueden realizarse tanto individualmente con las pistas como realizando reproducciones en vivo conjuntas que podrá oír como único consumidor final el director. Así, podrá realizar las correcciones e indicaciones que crea necesarias a cada integrante, pudiendo ejecutar las tareas del director en el ensayo. Al haber propuesto algunos segundos de demora, y no varios minutos, es posible incluso que mantenga una conversación en otro medio, en el cual no se incluyan sonidos del resto de integrantes, para poder frenar las ejecuciones cuando lo desea. Lo único que se perderá en cuestión de tiempos es el correspondiente a la suma de buffers, pudiendo frenar la ejecución esa cantidad de tiempo posterior a lo que cree que lo hace.

Una transmisión pública comienza algunos minutos antes de realmente hacerse pública, en este momento, inician las pruebas de entorno. Cumplida su tarea, se envía la señal de comienzo a todos los proveedores que, siguiendo la pista que se reproducirá automáticamente, comienzan su interpretación, generando el contenido. En el mismo, se graban las marcas temporales idénticas a las que se están reproduciendo como base. El contenido se guarda en un buffer de salida y, al completar un segmento, es solicitado por el servidor. Pese a ser solicitado seguirá guardándose en buffer por un tiempo, pudiendo volver a ser solicitado. El servidor solicitará nuevos segmentos luego de transcurrido el tiempo de reproducción de cada uno, que coincide con el tiempo de grabación, por lo cual ya habrá uno nuevo disponible. En caso de no

---

recibir a tiempo, comienza la tarea de nuevas solicitudes, en algunos casos, cambiando la o las resoluciones de ellas.

Al completarse el buffer del servidor, se brinda la disponibilidad del contenido a los consumidores finales, quienes pueden solicitar tantos segmentos como deseen, mientras estén disponibles en el servidor. El servidor ha de anunciar, al poner disponibles los segmentos, la duración del buffer. El consumidor utilizará un buffer idéntico al informado, con la salvedad de que el servidor debe conservar en un buffer negativo los segmentos que ya hayan transcurrido. El consumidor puede volver a solicitar dichos segmentos y es por esto que no deben eliminarse rápidamente. La lógica de solicitudes del consumidor es similar a la del servidor, a menos que éste decida manualmente recibir una resolución determinada.

Cuando el consumidor realiza una solicitud informa también el tipo de vista que posee. En caso de estar, por ejemplo, viendo la imagen del video de solo un integrante, debe responderse con ese segmento de video solo, pero con todos los segmentos de audio. No se provee la posibilidad de anular el audio de uno o algunos integrantes, aunque podría llegar a ser una extensión para las vistas de usuario.

### **Recomendaciones Extra**

La instancia de pruebas de entorno, o asimismo los ensayos, pueden servir para que el director de la orquesta regule el volumen de cada integrante. Esto, no solo en la práctica manual de su interpretación, sino, dándole una opción en el software para hacerlo. Es importante que cada proveedor realice la transmisión con el mismo hardware que haya realizado los ensayos. También debe mantener la relación de espacios físicos, dentro de lo posible. En caso de oír por parlantes lejanos en lugar de auriculares alámbricos, por ejemplo, estará acostumbrado a ejecutar sus obras con el retraso determinado por dicho hardware. Esto habrá producido ajustes en los ensayos que serían en vano si decidiera utilizar luego auriculares alámbricos. Lo mismo sucede si se utilizan auriculares inalámbricos en un caso y alámbricos en otro, teniendo en cuenta que poseen distintos retrasos en la entrega del sonido. Como recomendación general, se debe mantener el mismo lugar y equipamiento. Como indica Martínez Pegalajar, “lo ideal será que la sala de ensayo sea la misma que la del concierto. En caso contrario, hay que buscar unas condiciones de trabajo óptimas, tanto a nivel acústico como a nivel de confortabilidad” (2019, pág. 69).

---

Como ya ha sido mencionado en las páginas 63, 69 y 72, existe la posibilidad de tercerizar la tarea de transcodificación. Esto debe ser analizado dependiendo la oferta y demanda que posea nuestro hardware, y los tiempos de respuesta que pueda o no adicional esta tercerización. Es claro que el agregar eslabones a una cadena de envío agrega instancias en las cuales existen tiempos de retraso, no obstante, estos tiempos podrían ser, en algunos casos, menores a la reducción que producen a la demora de otros eslabones. De esta forma, puede evaluarse entonces la tercerización también de otra tarea, como lo es la distribución del contenido a los consumidores finales, o hasta podría dividirse la tarea de recolección entre más de un servidor, ubicados en diferentes sitios, si la arquitectura es utilizada de forma mundialmente distribuida. Recomendamos evaluar la posibilidad de derivar las tareas de transcodificación, beneficiándose de la computación en la nube. Adicionalmente, recomendamos la investigación de balanceo de cargas, aplicable a la tarea de distribución de contenidos. Esto es, no centralizar a todos los consumidores hacia el mismo servidor, si no, al más conveniente de más de uno disponible.

El balanceo de cargas podría ser aplicado también en la recolección del contenido, en caso de tener la transcodificación tercerizada y distribución balanceada. De esta forma, cada proveedor enviaría sus flujos al servidor más conveniente, y este, los reenviaría al servidor de transcodificación, o servidores de distribución, según corresponda. Sin embargo, creemos que este balanceo podría no resultar tan conveniente como las otras dos acciones que mencionamos, al menos cuando ellas no sean requeridas. Se convierte entonces, el balanceo de cargas aplicado a la recolección y distribución de contenido audiovisual de múltiples usuarios a sincronizar, una línea futura de investigación. Deseamos aclarar que la adopción de este balanceo puede motivarse tanto por una excesiva demanda de las capacidades de un servidor, como por una conveniencia geográfica que acorte los tiempos de transmisión, o incluso el conjunto de ellas.

En cuanto a la duración del buffer planteada, existe la posibilidad de que resulte de mayor interés, a los usuarios finales, reducir la cantidad de momentos donde no se tenga disponibilidad de los videos en alta calidad de todos los músicos. Esto, resignando un tiempo adicional en concepto de buffer al propuesto en este trabajo. Si ocurre de esta manera, puede prolongarse en tiempo de espera y/o cantidad de pedidos para obtener finalmente la calidad deseada. Posiblemente, con un buffer más largo, existan casos donde demoras momentáneas sean subsanadas con una mejora, también momentánea, del mismo entorno de red, dando lugar a

---

especular con la recepción correcta, sin riesgos, durante un tiempo mayor. De todas maneras, el ajuste dinámico de cantidad y calidad de contenido que se solicita al proveedor debe seguir existiendo, ya que el buffer puede ser más amplio al que hemos recomendado, pero no se alargará indefinidamente.

### **Conclusiones**

No existen diferencias absolutas entre la percepción de sincronización de un humano y la capacidad de ejecución musical sincronizada de otro, pese a haber ejercitado esta tarea. Esto radica, en principio, en que existe una habilidad innata que podría ser más precisa en quién aun no lo haya entrenado que en quien si lo haya hecho, sin importar cuando haya logrado agudizarla. En contraposición, también es posible que un músico inexperto posea un grado más privilegiado de esta habilidad que un oyente ejercitado en el tema. Adicionalmente, la tarea del músico se ve complejizada al sumar a su percepción una acción motriz que debe ser igualmente sincronizada, y lo cuál tampoco sería posible sin una buena anticipación de parte de él. Todo esto colabora a que las posibilidades de percibir errores sean más grandes que las de no cometerlos.

Realizar la ejecución siguiendo una pista base es un hecho familiar para los músicos de grupos de varias personas, quienes suelen seguir el ritmo de un guía, sea instrumental o gesticulado. Como se ha comentado, dicha pista será descargada con anterioridad a las transmisiones, para reproducirla directamente desde el equipo local de cada usuario, evitando demoras y fluctuaciones. El hecho de abstraer a un músico de la adición de retrasos en la entrega del sonido, al cual debe sincronizar su ejecución musical, le permite el ajuste rítmico de su propia ejecución. Esto evita los desfasajes progresivos que podrían alejar, o acercar, cada vez más una ejecución de la siguiente.

La posibilidad de negociación temporal entre participantes no es aplicable a los conjuntos guiados. Esto es independiente de la utilización de la pista propuesta, de hecho, se da en la presencialidad. La misma se presenta solo cuando no existe una referencia temporal a la cuál seguir. Esto sucede entre los acompañantes musicales y bailarines sin perder de vista que son solo dos personas intentando sincronizarse el uno con el otro, sin que la responsabilidad temporal sea completamente establecida en uno de ellos.

Replicar en los flujos de salida las marcas temporales de la guía, que se está reproduciendo mientras se realiza una ejecución en sincronía, garantiza una consecuente sincronización entre paquetes (Inter-Bundle), sin que deban adicionarse esfuerzos para lograrla.

---

Esto es, la sincronización entre distintos participantes de la ejecución conjunta. Adicionalmente, disminuye los efectos contrarios que podría ocasionar la deriva del reloj local de cada músico, ya que esta afecta de igual manera a la reproducción del contenido localmente como a su grabación.

Los efectos negativos que podría ocasionar un bajón en el flujo de bits de uno, o varios, proveedores de contenido, son minimizados por la gestión de solicitudes de paquetes en tiempo real, ejercida por el servidor. Esto gracias al ajuste dinámico de la cantidad y calidad de flujos a solicitar a cada uno de ellos. De igual manera, la posibilidad de afecciones al consumidor final también se reduce al solicitar solo los flujos necesarios. No todos los consumidores del contenido necesitarán recibir imágenes en alta definición de todos los proveedores. Se mantiene la posibilidad de hacerlo, pero a necesidad es determinada por las opciones de visualización.

En base a la investigación y propuesta realizada, podemos reafirmar la hipótesis planteada en la página 12. La transmisión en línea de contenidos audiovisuales sincronizados, emitidos por distintas personas, desde diversos lugares físicos, puede realizarse sin diferirlos considerablemente en el tiempo. El hecho de diferir, al menos unos pocos segundos, la transmisión, resulta necesario e inevitable para mantener la calidad de servicio deseada. En caso de no hacerlo, esta se vería degradada, bien por la omisión de sonidos, o por la manipulación de sus tiempos de reproducción. Si bien planteamos que no es necesario una gran cantidad de tiempo de retraso, este factor es considerado igualmente como una limitación. De hecho, no solo lo es la fluctuación de retardos, sino así también la demora de extremo a extremo inherente a la infraestructura de redes que se utilice. Los tiempos de retraso de extremo a extremo planteados en este trabajo son despreciables para los objetivos buscados.

A partir del desarrollo de una aplicación basada en la propuesta realizada pueden comenzar a realizarse transmisiones de obras musicalizadas sin la necesidad de recopilar, editar y unificar los distintos videos de los participantes. Adicionalmente, permite el hacerlo sin la necesidad de grabar anticipadamente cada una de las contribuciones. Esta nueva metodología no solo brinda comodidad y facilidades a quienes deseen realizar estas transmisiones, sino, una experiencia más realista, desconocida hasta el momento por los espectadores.

El resultado final de esta investigación se corresponde con el objetivo general planteado, habiendo diseñado la arquitectura de sincronización automática buscada. Esto, fue posible gracias al seguimiento individual de los objetivos específicos, partiendo de aquellos que permitiesen abordar las temáticas necesarias para generar conocimiento aplicable a los que

---

apuntaban de una forma más práctica a la solución del problema, representada por el objetivo general.

### **Líneas Futuras de Investigación**

Durante la confección del presente trabajo se tuvo en cuenta, como eje principal de los objetivos, la sincronización de los archivos multimedia como punto intermedio entre la recolección y la distribución del contenido. Los métodos de obtención han decantado como una necesidad para su desarrollo, planteando de esta forma las pruebas de entorno. Sin embargo, es posible extender la investigación en cuanto a ese aspecto de la arquitectura, aun sin despreciar el aporte mencionado. El caso de los métodos de distribución no ha sido explotado, con lo cual, abren una clara línea para investigaciones futuras. Existen mecanismos que podrían ser aplicados, previa investigación y análisis, para lograr, en algunos casos, acortar los tiempos de transmisión, o en otros quizá aumentarlos, pero brindando una mayor escalabilidad. En este sentido, la aplicación de balanceo de cargas, tanto en la recolección, como en la distribución, de contenido audiovisual de múltiples usuarios a sincronizar, representa una interesante extensión de nuestra investigación.

Una tarea que podría llegar a ser compleja por su demanda de procesos a ejecutar es la transcodificación de videos. En principio, pretendemos que sea realizada en el proveedor, y finalmente, también planteamos la posibilidad de que la realice el cliente, pero no podemos desligar al servidor de esta función. En casos donde sea necesario, es posible que esta tarea sea tercerizada, lo que podría, o no, resultar ventajoso también para las demoras de extremo a extremo. Esto merece un profundo análisis de los pros, las contras, y las situaciones donde podría o no resultar ventajosa. Se desprende así la necesidad de analizar las posibilidades de tercerización de transcodificación en una arquitectura de sincronización multiusuario para transmisión en directo de obras musicalizadas.

Como ya hemos mencionado en la página 58, existe la posibilidad de aplicar aprendizaje profundo al proceso de sincronización de los archivos multimedia. Si bien, en esa misma página, no lo recomendamos, por desviarse de los objetivos buscados, es una herramienta a tener en cuenta para otros aspectos. La estimación de los tiempos de buffer eficientes, dado un conjunto de proveedores y/o clientes, teniendo en cuenta sus infraestructuras de redes, es un proceso que puede nutrirse de este tipo de prácticas. Recomendamos ahondar a futuro en su investigación aplicada a este aspecto.

---

**Acrónimos****WAN**            *Wide Area Network*

Red de área amplia, con alcance mayor a redes locales o metropolitanas, que las interconecta. Utilizado para referirse a un alcance mundial.

**4D**            *Cuadridimensional*

Utilizado para referirse a audiovisuales de imágenes en 3 dimensiones, a los cuales se agregan otros estímulos sensoriales, normalmente por medio de estímulos con viento, agua y olores. Frecuentemente utilizado en el ambiente cinematográfico.

**3D**            *Tridimensional*

Utilizado para referirse a audiovisuales de imágenes en 3 dimensiones. Frecuentemente utilizado en el ambiente cinematográfico. Suele ser necesario utilizar anteojos diseñados para visualizar la imagen de forma correcta, sin percibirla como dos imágenes similares superpuestas.

**FPS**            *Frames per Second*

Utilizado para referirse a la cantidad de imágenes que se visualizan en un segundo de video. Traducido al castellano como fotos por segundo o cuadros por segundo.

**IDES**            *Inter Device Synchronization*

En castellano, sincronización entre dispositivos. Suele utilizarse para referirse a dispositivos que se encuentran en un mismo ámbito o lugar físico.

**IDMS**            *Inter-Destination Multimedia Synchronization*

En castellano, sincronización multimedia entre destinatarios. Suele utilizarse para referirse a dispositivos pertenecientes a diferentes ámbitos o lugares físicos.

**UM**            *Unidades de medios*

Se utiliza para referirse a la fracción del contenido multimedia que se envía en conjunto.

**QoE**            *Quality of experience*

Traducido en castellano como calidad de experiencia.

**NTP**            *Network Time Protocol*

En castellano, protocolo de tiempo de red de trabajo. Utilizado para la sincronización de relojes.

**PTP**            *Precision Time Protocol*

En castellano, protocolo de tiempo de precisión. Utilizado para la sincronización de relojes.

---

**GPS**            *Global Positioning System*

Traducido en castellano como sistema de posicionamiento global. Se utiliza para conocer la ubicación geográfica.

**USB**            *Universal Serial Bus*

Traducido en castellano como bus universal en serie. Es uno de los puertos más conocidos en las computadoras. Transmite bits en serie.

**QoS**            *Quality of Service*

Traducido en castellano como calidad de servicio.

**ms**              *Milisegundo*

Unidad de medida temporal que representa la milésima parte de un segundo.

**AMD**            *Acompañante Musical de Danza*

Se refiere al músico que ejecuta una obra conjunta con un bailarín.

**cm**              *Centímetro*

Unidad de medida que representa la centésima parte de un metro.

**DAC**            *Digital to Analogue Converter*

En castellano, Convertidor Digital a Analógico. Se utiliza para la conversión de señales.

**4K**              *Cuatro Mil (Píxeles)*

Resolución de pantalla con aproximadamente 4000 píxeles en horizontal.

**UHD**            *Ultra Alta Definición*

Resolución de pantalla que se corresponde con la llamada 4k, con sus 3.840 x 2.160 píxeles.

**HDR**            *Alto Rango Dinámico*

Cualidad de pantallas que permite una alta diferencia entre sus tonos más iluminados con los menos iluminados.

**p**                *Píxeles*

Se utiliza posterior a la cantidad de píxeles que se expresan en las distintas resoluciones.

**Hz**              *Hercio*

Es una unidad de frecuencia, utilizada en pantallas para expresar la frecuencia con la que se refresca la imagen.

**LCD**            *Liquid Crystal Display*

En castellano, pantalla de cristal líquido.

---

**Kbps**      *Kilobit por Segundo*

Unidad de medida de transferencia de datos que representa 1000 (mil) bits por segundo.

**Mbps**      *Megabit por Segundo*

Unidad de medida de transferencia de datos que representa 1000 (mil) kilobits por segundo.

**Full HD**      *Alta Definición Total*

Resolución de pantalla de 1920×1080p.

**HD**      *Alta Definición*

Resolución de pantalla de 1280x720p.

**nHD**      *Noveno de Alta Definición*

Resolución de pantalla que representa la novena parte de Full HD, 640x360p.

---

## Referencias

- Advanced Television Systems Committee. (26 de Junio de 2003). Relative Timing of Sound and Vision for Broadcast. *Doc. IS-191*. Washington D. C.
- Aorus. (13 de Agosto de 2022). <https://www.aorus.com/es-ar/monitors/AORUS-FI27Q-X/Specification>
- Batteram, H., Damm, G., Mukhopadhyay, A., Philippart, L., Odysseos, R., & Urrutia-Valdés, C. (2010). Delivering Quality of Experience in Multimedia Networks. *Bell Labs Technical Journal*, 175-193.
- BenQ. (13 de Agosto de 2022). <https://www.benq.com/en-us/projector/gaming-projector/tk700sti/specifications.html>
- Bleumers, L., Wallendael, G., De Cock, J., Geeraert, K., Vercammen, N., Van den Broeck, W., . . . Demeester, P. (2012). Assessing the importance of audio/video synchronization for simultaneous translation of video sequences. En *Multimedia Systems* (Vol. 18, págs. 445-457). Springer.
- Chuang, P.-H. (2005). The conductor and the ensemble: From a psychological aspect.
- González Lapuente, A. (2003). *Diccionario de la música*. Madrid: Alianza Editorial.
- He, L., Cai, P., & Zhau, J. (2009). CS-CS stream media low-level synchronization control mechanism. *2009 ISECS International Colloquium on Computing, Communication, Control, and Management*, (págs. 411-414).
- Huang, Z., & Nahrstedt, K. (2013). Evolution of Temporal Multimedia Synchronization Principles: A Historical Viewpoint. *ACM Transactions on Multimedia Computing, Communications and Applications*.
- Ibarrola, E., liberal, F., Taboada, I., & Ortega, R. (2009). Web QoE Evaluation in Multi-Agent Networks: Validation of ITU-T G.1030. *2009 Fifth International Conference on Autonomic and Autonomous Systems*, (págs. 289-294).
- Kopiez, R. (2002). Making music and making sense through music. *The new handbook of research on music teaching and learning: A project of the Music Educators National Conference* (págs. 522-541). Oxford University Press.
- Laguna, A., & Shifres, F. (2012). Indicios Visuales y Auditivos en el Ajuste Sincrónico del Pulso Subyacente Entre Bailarines y Acompañantes Musicales. *Congreso de la Sociedad Ibérica de Etnomusicología*. Lisboa: Sociedad Iberica de Etnomusicología.

- 
- LG. (13 de Agosto de 2022). <https://www.lg.com/ar/monitores/lg-19M38A-B>
- Liu, H., & Zarki, M. (2003). *Towards the Delay and Synchronization Control for Networked Real-Time Multi-Object Multimedia Applications*. Irvine, California.
- Liu, H., & Zarki, M. (2005). *Adaptive Delay and Synchronization Control for Wi-Fi Based Mobile AV Conferencing*. Irvine, California.
- Liu, H., & Zarki, M. (2006). *An adaptive delay and synchronization control scheme for Wi-Fi based audio/video conferencing*. Irvine, California: Springer.
- Liu, H., & Zarki, M. (2010). *Adaptive Delay and Synchronization Control for Wi-Fi Based AV Conferencing*.
- Liu, H., & Zarki, M. E. (2003). *A Synchronization Control Scheme for Real-Time Streaming Multimedia Applications*. Irvine, California.
- Malbrán, S. (2007). *Sincronía Rítmica y Tempo: Un Estudio con Adultos Músicos*.
- Mark E. (2 de Agosto de 2022). Comunicación personal.
- Martínez Pegalajar, P. (2019). El director de orquesta en el ensayo: análisis teórico y práctico. *Revista del Real Conservatorio Superior de Música de Madrid*(26), 179-201.
- Mata Galiano, V. (2014). *Análisis y comparativa de los protocolos de transmisión de vídeo adaptativo por internet*.
- Montagud, M., Boronat, F., Martínez, M., Belda, J., & Cesar, P. (2015). Impacto de Parámetros de QoS en Aspectos de QoE: Análisis desde el Punto de Vista de la Sincronización Multimedia. *Jornadas de Ingeniería Telemática - JITEL 2015* (págs. 355-362). Gandia, Valencia.: CWI's Institutional Repository.
- Montagud, M., Cesar, P., Boronat, F., & Jansen, J. (2018). *MediaSync*. Darmstadt: Springer.
- Mued, L., Lines, B., Furnell, S., & Reynolds, P. (2003). The effects of audio and video correlation and lip synchronization. *Campus-Wide Information Systems*, 20, 159-166.
- Nilsson, E. (2018). *Evaluation of how clock synchronisation protocols affects inter-sender synchronisation of live continuous multimedia*. Tesis de maestría, Umeå University, Engineering and Technology, Umeå. <http://umu.diva-portal.org/>
- Pena, J., Anglés, H., & Querol, M. (1954). *Diccionario de la música Labor*. Barcelona: Labor.
- Popoca, J. (2016). Un modelo para la música: teoría del ritmo. *Metatheoria—Revista de Filosofía e Historia de la Ciencia*, 6(1), 73-78.

- 
- Previtali, F. (1969). *Guía para el estudio de la dirección orquestal*. Buenos Aires: Ricordi Americana.
- Quintero, E. (1 de Agosto de 2022). Comunicación personal.
- Retting, F. (1990). Director de orquesta. (I. Molinares, Entrevistador)
- Román Sarmiento, R. (2020). Detección de Sincronización Audiovisual Utilizando Deep Learning.
- sagora. (2021). <https://sagora.org/>
- Saini, M. K., & Ooi, W. T. (2018). Automated Video Mashups: Research and Challenges. En M. Montagud, P. Cesar, F. Boronat, & J. Jansen, *MediaSync* (págs. 167-190). Darmstadt: Springer.
- Sáliche, L., & Rodríguez, M. (15 de agosto de 2021). Ensayos solitarios, distancia social y protocolos: ¿cómo sobreviven las orquestas y los coros en pandemia? *Infobae*. <https://www.infobae.com>
- Salmon, R., & Mason, A. (2008). Factors affecting perception of audio-video synchronisation in television.
- Shifres, F., & Holguín Tovar, P. (Marzo de 2015). El Desarrollo de las Habilidades Auditivas de los Músicos. Teoría e Investigación. La Plata: GITEV.
- Staelens, N., De Meulenaere, J., Bleumers, L., Van Wallendael, G., De Cock, J., Geeraert, K., . . . Demeester, P. (2012). Assessing the importance of audio/video synchronization for simultaneous translation of video sequences. (M. Angelides, Ed.) *Multimedia Systems*, 18, págs. 445-457.
- Steinberg, G. (1946). Prejuicios sobre la interpretación musical. *Revista Musical Chilena*, 2(13), 13-18.
- Steinmetz, R. (1996). Human Perception of Jitter and Media Synchronization. *IEEE Journal on Selected Areas in Communications*, Vol. 14, No. 1, January 1996, 61-72.
- Steinmetz, R., & Engler, C. (1993). Human Perception of Media Synchronization.
- Unión Internacional de Telecomunicaciones. (Noviembre de 1998). Temporización Relativa del Sonido y la Imagen Para la Radiodifusión. *RECOMENDACIÓN UIT-R BT.1359-1*. <https://www.itu.int/rec/R-REC-BT.1359/es>
- Valencia Restrepo, D. (2001). Beethoven y el Metrónomo. *Revista Universidad de Antioquia*(265).

- 
- van Deventer, M., Stokking, H., Niamut, O., Walraven, F., & Klos, V. (2008). Advanced Interactive Television Services Require Content Synchronization. *In 2008 15th International Conference on Systems, Signals and Image Processing* (págs. 109-112). IEEE.
- Villarreal Rodríguez, G. (2016). Evolución y desarrollo de la dirección orquestal en México y el mundo. *RICSH Revista Iberoamericana de las Ciencias Sociales y Humanísticas*, 5(9).
- Vorterix. (19 de Septiembre de 2022). *Vorterix*. <https://vorterix.com>
- Yilmaz, S., Tekalp, A. M., & Unluturk, B. D. (2015). Video Streaming Over Software Defined Networks with Server Load Balancing. *2015 International Conference on Computing, Networking and Communications (ICNC)*, (págs. 722-726). Sariyer. doi:10.1109/ICCNC.2015.7069435
- Zarki, M., Cheng, L., Liu, H., & Wei, X. (2003). An Interactive Object Based Multimedia System for IP Networks. Irvine.