UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Gaona Romero, Hugo. *Desenvolupament d'un sistema de sincronització automàtica entre fonts d'àudio locals i senyals de televisió estrangeres.*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Enginyeria de Dades)

This version is available at https://ddd.uab.cat/record/337323

under the terms of the CC BY-SA license.

Desenvolupament d’un sistema de sincronització automàtica entre fonts d’àudio locals i senyals de televisió estrangeres

Hugo Gaona Romero

Grau en Enginyeria de Dades — Universitat Autònoma de Barcelona

Tutor: Víctor Montilla | hugo.gaona@autonoma.cat | Juny 2026

Resum — Les emissions esportives en directe arriben al consumidor a través de fonts tecnològicament heterogènies —satèl·lit, TDT, ràdio per internet— que introdueixen retards temporals respecte de l'event real, variables entre 0,5 i 30 s segons la font i l'equip de processament digital. Quan un espectador combina dues d'aquestes fonts simultàniament, per exemple seguint la imatge d'un canal de televisió amb la locució d'una emissora de ràdio en el seu idioma nadiu o locutor preferit, l'asincronisme és evidentment perceptible i degrada l'experiència. Avui dia no existeix cap solució automàtica accessible per a l'usuari domèstic. Aquest treball proposa, dissenya i avalua experimentalment un sistema de sincronització automàtica d'àudio capaç de detectar i corregir aquest retard sense cap intervenció manual, operant sobre les dues fonts d'àudio en temps real. El sistema implementa un pipeline de tres etapes — filtratge de soroll, extracció de característiques i detecció del retard— i compara quatre combinacions d'algorismes sobre set gravacions simultànies reals de la final de la Copa del Rei 2026 (Atlètic de Madrid vs Reial Societat). Els resultats mostren que la combinació Bandpass+Onset+DTW ofereix la major estabilitat en àudio real (std=0,17 s, 1 resincronització/22 finestres) amb 22,8 ms/finestra. El profiling real amb cProfile identifica el filtre Wiener com a coll d'ampolla (58% del temps total).

Paraules clau — sincronització d'àudio, onset detection, GCC-PHAT, dynamic time warping, MFCC, processament de senyal en temps real (RT computing), hardware encastat, emissió esportiva en directe.

Abstract — Live sports broadcasts reach consumers through technologically heterogeneous sources —satellite, DTT, internet radio— introducing unpredictable time delays between 0.5 and 30 seconds. When a viewer simultaneously combines two of these sources, the asynchrony is clearly perceptible and degrades the experience. No automatic solution currently exists for the domestic user. This paper proposes, designs and experimentally evaluates an automatic audio synchronization system capable of detecting and correcting this delay without manual intervention, operating on both audio streams in real time. The system implements a three-stage pipeline —noise filtering, feature extraction and delay detection— and compares four algorithm combinations over seven simultaneous real recordings of the 2026 Copa del Rey final. Bandpass+Onset+DTW achieves the highest stability on real heterogeneous audio (std=0.17 s, 1 resync/22 windows) at 22.8 ms per 10-second window. Real cProfile profiling identifies the Wiener filter as the main bottleneck (58% of total time).

Index Terms — audio synchronization, onset detection, GCC-PHAT, dynamic time warping, MFCC, real-time signal processing, embedded hardware, live sports broadcast.

1 Introducció

La televisió i la ràdio esportiva s'han fragmentat en un ecosistema de fonts heterogènies: TDT, satèl·lit, IPTV, plataformes de streaming i emissores de ràdio per internet. Cadascuna introdueix una latència de transmissió diferent: la TDT acumula menys de 10 s degut a la compressió MPEG i la multiplexació, mentre que les plataformes d'streaming (CDN, recodificació, HLS/DASH) poden arribar als 30 s o més [1].

Això genera situacions molt concretes i quotidianes: el locutor de ràdio narra un gol mentre la televisió encara mostra la jugada anterior; les notificacions mòbils anuncien el resultat d'una acció que la televisió no ha emès encara; o l'espectador estranger que combina imatges d'un canal internacional amb la locució local no pot sincronitzar-les manualment de forma precisa.

Avui dia, l'única solució és l'ajust manual per assaig i error, imprecís i repetitiu en cada sessió. Aquest treball automatitza completament la detecció i correcció d'aquest retard, sense cap referència externa, en temps real i sense intervenció de l'usuari.

Contribucions: (1) implementació i comparació de quatre combinacions d'algorismes; (2) validació sobre set fonts simultànies reals; (3) profiling computacional real amb cProfile identificant colls d'ampolla; (4) anàlisi de portabilitat en hardware encastat de baix cost.

L'estructura de l'article segueix el flux metodològic descrit: context i treballs relacionats (Secció 2), metodologia i arquitectura (Seccions 3-4), resultats i discussió (Seccions 5-6) i conclusions (Secció 7).

2 Context i treballs relacionats

El problema de l'estimació del retard entre senyals d'àudio és clàssic. Knapp i Carter [2] van proposar la Generalized Cross-Correlation with Phase Transform (GCC-PHAT) el 1976, correlació creuada normalitzada en fase que millora la robustesa al soroll. Wang [3] va introduir l'Audio Fingerprinting de Shazam (2003), que identifica fragments independentment del retard i la degradació. Els Mel-Frequency Cepstral Coefficients (MFCC) [4] de Davis i Mermelstein (1980) són l'estàndard en reconeixement acústic gràcies al baix cost. El Dynamic Time Warping (DTW) [5] de Sakoe i Chiba (1978) alinea seqüències temporals amb variació de velocitat. La detecció de transitoris sonors ha estat estudiada per Bello, Daudet, etc. [6].

El filtre de Wiener [7] elimina el soroll de fons estimant el seu perfil espectral i restant-lo del senyal. La Voice Activity Detection (VAD) [10] detecta i emmudeix els segments on no hi ha activitat sonora. Totes dues tècniques es combinen amb el filtre Bandpass per millorar la qualitat del senyal abans de l'extracció de característiques. La biblioteca librosa [8] proporciona les implementacions en Python utilitzades en aquest treball.

A diferència dels treballs anteriors, el nostre cas d'ús opera sobre fonts completament heterogènies —una emissora de ràdio per internet i un canal de televisió per satèl·lit— en temps real i sense intervenció humana. La diferent textura acústica entre fonts (codificadors, locutors, entorns sonors) fa que la GCC-PHAT clàssica trobi ambigüitats significatives, limitació que aquest treball quantifica i resol experimentalment, obtenint elements clau de comparació entre fonts de transmissió heterogènies.

3 Metodologia

Aquesta secció descriu el mètode emprat per resoldre el problema de la sincronització automàtica. El mètode segueix un full de ruta de quatre fases: (1) exploració de les tècniques disponibles en la literatura, (2) selecció basada en criteris de robustesa, cost computacional i portabilitat, (3) implementació de l'arquitectura en pipeline de tres etapes, i (4) validació sota dues condicions complementàries.

La validació utilitza el mètode Ground Truth com a primera fase: s'aplica un retard artificial i conegut a un àudio real i es mesura quina combinació el detecta amb més precisió. Això permet tenir una referència objectiva de la capacitat de cada algorisme en condicions controlades. La segona fase usa àudio real de fonts heterogènies, on el retard real és desconegut, i avalua l'estabilitat del sistema entre finestres consecutives.

Aquesta metodologia es concreta en els punts clau següents: dades experimentals (3.1), pipeline de sincronització (3.2), finestra lliscant (3.3), correlació de fonts (3.4), algoritmes seleccionats (3.5) i proposta d'arquitectura (Secció 4). Les combinacions avaluades i l'entorn de test es descriuen a la Secció 4.4.

3.1 Dades experimentals: set fonts simultànies

El corpus està format per set gravacions simultànies de la final de la Copa del Rei 2026. De cada font s'ha extret un segment de 120 s en WAV monocanal a 16.000 Hz i 16 bits (Taula I). El 16 kHz cobreix fins a 8 kHz per Nyquist, suficient per a veu i sons esportius, amb la meitat del cost computacional de 32 kHz.

TAULA I. SET FONTS ENREGISTRADES DE LA FINAL COPA DEL REI 2026

Font	Tecnologia	Còdec àudio	Format
La 1 HD	Satèl·lit/TDT	AC-3 5.1 / 48 kHz	.ts
La 1 UHD	Satèl·lit/TDT	AC-3 5.1 / 48 kHz	.ts
TV3	Satèl·lit/TDT	AC-3 5.1 / 48 kHz	.ts
COPE	Ràdio internet	AAC-LC / 44,1 kHz	.mp4
RAC1	Ràdio internet	AAC-LC / 44,1 kHz	.mp4
RNE	Ràdio internet	AAC-LC / 44,1 kHz	.mp4
SER	Ràdio internet	AAC-LC / 44,1 kHz	.mp4

Limitació important: els arxius es van gravar de forma independent. El tall al segon 120 no garanteix el mateix instant real del partit en tots els clips, de manera que el retard mesurat en parelles reals inclou el retard tecnològic més la diferència de punt d'inici d'enregistrament. Per això, a les parelles reals mesurem l'estabilitat del sistema (consistència entre finestres) i no l'error absolut.

3.2 El pipeline de sincronització

El sistema implementa un pipeline de tres etapes aplicat sobre finestres lliscants (Fig. 1). Cada etapa transforma progressivament el senyal: de les mostres d'àudio brutes fins a un únic número en segons que representa el retard estimat.

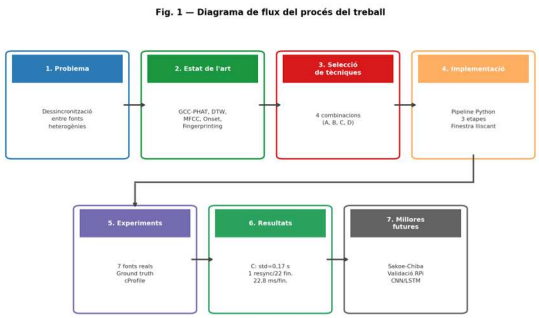


Fig. 1. Diagrama de flux del procés del treball. Les tres etapes s'executen sobre finestres lliscants solapades. La resincronització dinàmica s'activa cada pas si el retard canvia.

L'elecció de W=10 s s'ha obtingut per grid search (cerca exhaustiva sobre un conjunt de valors candidats per trobar el paràmetre òptim) sobre sis valors [3, 5, 7, 10, 12, 15 s]: és el millor compromís entre precisió (més mostres = millor estimació) i la latència de resposta del sistema.

3.3 Finestra lliscant: acumulació vs latència

Un concepte crític per entendre el sistema és la distinció entre la mida de la finestra i la latència de resposta. Són dues coses completament diferents.

La mida de la finestra W és el temps d'àudio que el sistema acumula per fer una estimació del retard prou fiable. Amb W=10 s el sistema disposa de prou transitoris sonors (xiulets, reaccions del públic) per trobar una alineació robusta.

La latència de resposta és el temps que tarda el sistema a detectar i corregir un canvi de retard. Depèn del pas entre finestres (5 s) i del temps de procés del pipeline (22,8 ms per a C). Per tant, la latència de resposta del sistema és de 5 s més 22,8 ms, no de 10 s.

En la pràctica, el retard entre fonts esportives és estacionari a curt termini (no canvia sobtadament en menys d'un minut), de manera que una latència de resposta de 5 s és perfectament acceptable per a l'ús domèstic previst.

3.4 Correlació de fonts

El procés de correlació és el cor del sistema. Imaginem dues tires de paper amb grups de punts (els transitoris sonors) distribuïts en el temps: una sèrie correspon a la ràdio i una altra a la televisió. La feina del sistema és trobar quant s'ha de desplaçar una de les tires perquè els patrons coincideixin al màxim.

La GCC-PHAT resol això en el domini freqüencial: multiplica els espectres de les dues seqüències i la transformada inversa de Fourier dona directament la funció de similitud per a tots els desplaçaments possibles alhora. És extremadament eficient (0,3–0,7 ms) però assumeix que el patró dels dos senyals és prou similar per tenir un màxim dominant, condició que falla en fonts molt heterogènies.

El DTW resol el mateix problema de forma totalment diferent: comprova explícitament totes les possibles alineacions entre frames consecutius i troba la seqüència de parelles (frame de ràdio, frame de TV) que minimitza el cost total d'alineament. És més lent (2,2–2,7 ms) però no requereix que el patró global sigui similar, només que els frames locals siguin comparables —condició molt més fàcil de complir entre fonts heterogènies. La Fig. 6 fa visible això.

En resum: GCC-PHAT és com buscar una agulla en un paller preguntant «on està la zona de màxima densitat d'agulles?»; DTW és com agafar-les una per una i aparellar-les pel cost mínim. El primer és més ràpid però confon zones de densitat similar; el segon és més robust però més costós.

3.5 Algoritmes i biblioteques seleccionats

Durant el treball s'han emprat principalment els algoritmes i biblioteques següents. La Taula II recull la seva funció, la configuració aplicada i la font d'implementació. Als sub-apartats 4.1, 4.2 i 4.3 es descriurà com s'integren en el pipeline.

TAULA II. ALGORITMES, CONFIGURACIÓ I FONTS D'IMPLEMENTACIÓ

Algoritme	Funció	Configuració clau	Font / Repositori
Onset Detection	Detecta transitoris sonors	hop_size=512, sr=16 kHz	librosa [8]
MFCC	13 coeficients acústics/frame	n_mfcc=13, hop_size=512	librosa [8]
GCC-PHAT	Correlació creuada en fase	FFT sobre W=10 s	scipy.signal
DTW	Alineació frame a frame	max_frames=300	Impl. pròpia (Python)
Bandpass IIR	Filtre passa-banda	100–7.900 Hz, ordre=4	scipy.signal
Wiener espectral	Subtracció soroll de fons	FFT per frame de 20 ms	scipy.signal
VAD	Emmudeix silencis	frame=20 ms, thr. energia	Impl. pròpia (Python)

Tots els algoritmes han estat implementats en Python 3 sobre un PC de sobretaula amb Windows. Les biblioteques scipy i librosa proporcionen les implementacions base per a Onset Detection, MFCC, GCC-PHAT, Bandpass i Wiener. DTW i VAD s'han implementat de forma pròpia seguint les especificacions dels articles originals [5, 10].

4 Proposta d'arquitectura del sistema

4.1 Etapa 1: Filtratge de soroll

L'Etapa 1 filtra freqüències irrellevants per a la detecció de transitoris. Per a cada combinació s'aplica una única opció de les tres disponibles:

Opció A — Bandpass IIR Butterworth ordre 4 (100–7.900 Hz, ~4,3 ms/finestra): elimina el ronc elèctric (50 Hz) i soroll d'alta freqüència irrellevant per a veu. Usat a les combinacions A i C.

Opció B — VAD+Bandpass: emmudeix silencis per energia de frame (20 ms) i aplica el Bandpass, millorant la qualitat per a MFCC. Usat a la combinació B.

Opció C — Wiener espectral (~50,7 ms/finestra): estima i resta el soroll de fons per FFT, identificat com a coll d'ampolla per cProfile. Usat a la combinació D.

4.2 Etapa 2: Extracció de característiques

L'Etapa 2 extreu les característiques de l'àudio filtrat. Dues opcions alternatives:

Opció A — Onset Detection [6]: genera un valor per frame de 32 ms representant l'energia dels transitoris. Els seus pics corresponen als instants d'esdeveniments rellevants comuns entre les dues fonts (xiulets, reaccions del públic). Les Figs. 2 i 3 mostren el problema i la detecció.

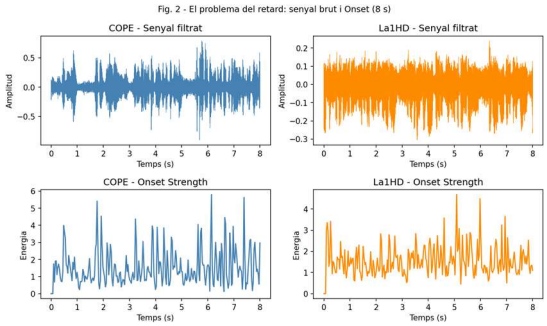


Fig. 2. El problema del retard: senyal brut (dalt) i Onset (baix) de COPE i La1HD (8 s). El desplaçament Δt entre els pics de les dues corbes és el retard a detectar i corregir.

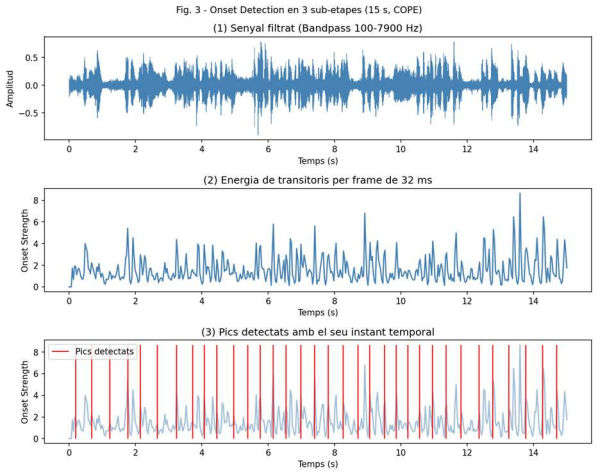


Fig. 3. Onset Detection en 3 sub-etapes (15 s, COPE): (1) senyal filtrat, (2) energia de transitoris per frame de 32 ms, (3) pics detectats amb el seu instant temporal.

Opció B — MFCC [4]: genera 13 coeficients per frame que descriuen l'empremta acústica completa, capturant la textura tonal. Més informatiu però més costós que l'Onset (Fig. 4).

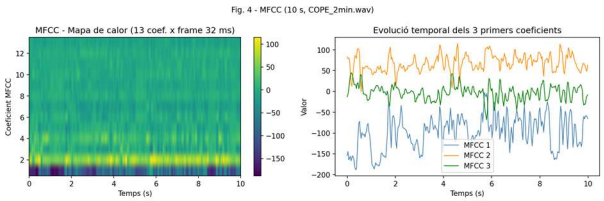


Fig. 4. MFCC (10 s, COPE_2min.wav): mapa de calor de 13 coeficients per frame de 32 ms (esquerra) i evolució temporal dels 3 primers coeficients (dreta).

4.3 Etapa 3: Detecció del retard

L'Etapa 3 pren les característiques de les dues fonts i calcula el retard. Dues opcions alternatives:

Opció A — GCC-PHAT [2]: correlació creuada normalitzada en fase. Eficient (0,3–0,7 ms) però sensible a l'ambigüitat de pic en fonts heterogènies (Fig. 5).

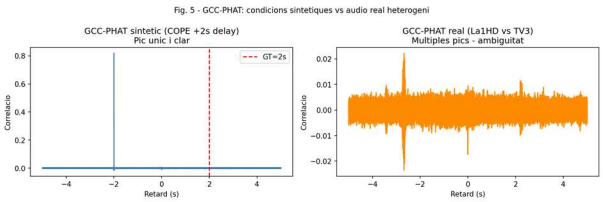
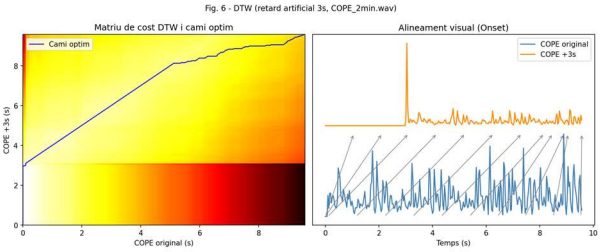


Fig. 5. GCC-PHAT: pic unívoc en condicions sintètiques (esquerra). En àudio real La1HD vs TV3 (dreta), múltiples pics de magnitud similar generen ambigüitat.

Opció B — DTW [5]: alinea frame a frame minimitzant el cost acumulat, robust davant heterogeneïtat. Cost 2,2–2,7 ms/finestra (màx. 300 frames). La Fig. 6 mostra la matriu de cost i l'alineament resultant.

Fig. 6. DTW (retard artificial 3 s, sobre COPE_2min.wav): matriu de cost i camí d'alineament òptim (esquerra); alineament visual de les



dues corbes Onset (dreta). Cada fletxa connecta instants equivalents.

4.4 Combinacions avaluades i entorn de test

Combinant les opcions de cada etapa s'obtenen fins a 8 combinacions possibles. S'han seleccionat les quatre que representen els extrems del trade-off cost–robustesa i permeten aïllar l'impacte de cada decisió de disseny:

TAULA III. LES QUATRE COMBINACIONS AVALUADES.

Comb.	Filtre	Feature	Detector	Justificació selecció
A	Bandpass	Onset	GCC-PHAT	Combinació mínima. Referència de velocitat.
B	VAD+BP	MFCC	GCC-PHAT	Millora filtre+feature. Ref. precisió sintètica.
C	Bandpass	Onset	DTW	Aïlla impacte del detector en àudio real.
D	Wiener	MFCC	DTW	Filtre+feature màxims. Ref. robustesa teòrica.

Les quatre combinacions restants (Bandpass+MFCC+GCC, VAD+Onset+GCC, Bandpass+MFCC+DTW i VAD+Onset+DTW) es descarten per evitar redundància: A i C ja cobreixen Onset, B i D ja cobreixen MFCC, i la comparació

A vs C i B vs D permet aïllar l'impacte del detector independentment de la feature.

Entorn de test: tots els experiments s'han executat en un PC de sobretaula amb Windows, usant Python 3 i les biblioteques de la Taula II. Les proves s'han llançat des de la línia de comandes (PowerShell). S'han usat dos tipus de tests: (1) retards artificials coneguts (ground truth), aplicant un desplaçament temporal conegut a COPE_2min.wav i comparant-lo amb la còpia original; (2) parelles d'àudio real de fonts heterogènies, on el retard real és desconegut.

5 Resultats

5.1 Corbes Onset: heterogeneïtat entre fonts

La Fig. 7 mostra les corbes d'Onset de les sis fonts durant 20 s. Les ràdios (COPE, SER, RNE, RAC1) presenten pics molt definits gràcies a la compressió d'àudio i la narració en primer pla, sense so d'estadi. Les televisions (La1HD, TV3) mostren una corba més densa i sorollosa, fruit de la barreja entre narració, ambient d'estadi i veus del públic. Això explica directament per què les parelles TV–ràdio són les més difícils per a GCC-PHAT.

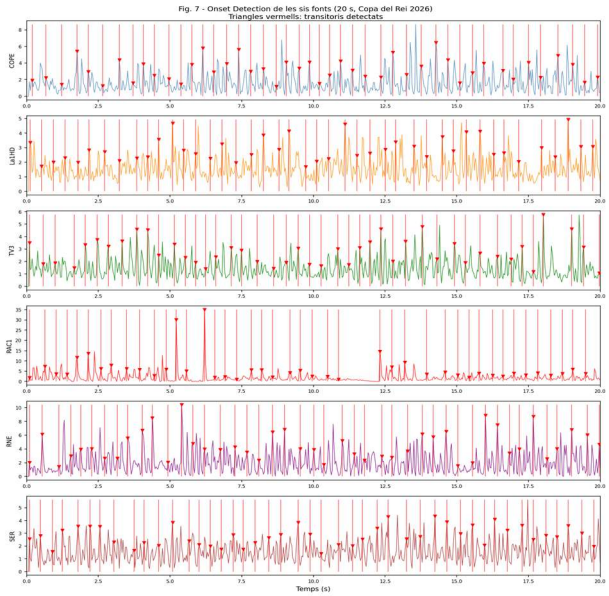


Fig. 7. Onset Detection de les sis fonts (20 s, Copa del Rei 2026). Triangles vermells: transitoris detectats. Ràdios: pics nets. TV: major densitat de soroll de fons.

Això té una implicació pràctica important: el tipus de transmissió influeix directament en la qualitat dels transitoris disponibles per a la correlació. Les parelles homogènies (ràdio–ràdio o TV–TV) comparteixen la mateixa textura de fons, cosa que facilita la correlació. Les parelles heterogènies (ràdio–TV) són el cas més exigent i el que justifica l'ús de DTW en comptes de GCC-PHAT.

5.2 Profiling computacional real (cProfile)

El profiling real (Fig. 8) identifica la distribució exacta del temps per etapa. El filtre Wiener és el coll d'ampolla de D: 50,7 ms dels 88 ms totals (58%), per les ~160 FFTs per finestra. El Wiener és costós perquè realitza una FFT per cada frame de 20 ms, resultant en ~160 FFTs per finestra de 10 s, cadascuna amb cost O(N log N). C és la més eficient: 29 ms, 4.122 crides vs 35.401 de D.

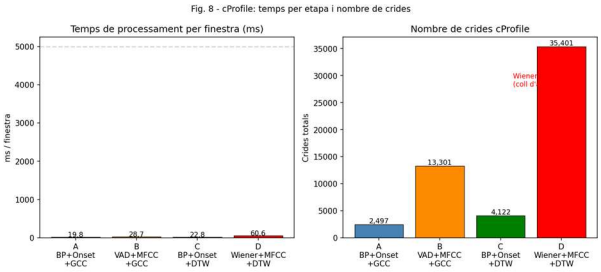


Fig. 8. cProfile: temps per etapa (esquerra) i nombre de crides (dreta). El Wiener de D és el coll d'ampolla (58%). C és la més eficient i senzilla.

5.3 Validació amb ground truth (N=5 averaging)

El ground truth és un mètode de validació on s'aplica un retard artificial conegut a un àudio (COPE_2min.wav) i es mesura amb quina precisió cada combinació el detecta. S'han testat retards de 2, 5, 10 i 15 s, repetint N=5 vegades per reduir la variabilitat d'inicialització de les biblioteques.

La Taula IV-A mostra el retard detectat per cada combinació. Valors propers al GT indiquen major precisió. Les combinacions C i D (DTW) retornen valors propers a 0 degut al problema estructural de traceback descrit a la Secció 6.

TAULA IV-A. RETARD DETECTAT (s) PER COMBINACIÓ I RETARD GT (N=5 AVERAGING).

GT (s)	Detectat A (s)	Detectat B (s)	Detectat C (s)	Detectat D (s)
2,0	4,62	2,51	0,16	0,35
5,0	3,87	2,51	0,13	0,24
10,0	8,34	5,01	0,03	0,06
15,0	0,00	15,01	0,00	0,00

La Taula IV-B mostra el temps de processament per finestra per cada combinació i retard GT.

TAULA IV-B. TEMPS DE PROCESSAMENT PER FINESTRA (ms) PER COMBINACIÓ I RETARD GT

GT (s)	T/fin. A (ms)	T/fin. B (ms)	T/fin. C (ms)	T/fin. D (ms)
2,0	28,7 *	16,5	16,2	33,6
5,0	28,7	16,2	12,9	32,4
10,0	23,4	29,7	12,8	61,1
15,0	12,3	42,5	13,0	32,1
Mitjana	23,3	26,2	13,7	39,8

(*) Artefacte d'inicialització de librosa en la primera execució. Valor real ~20-28 ms.

5.4 Validació en àudio real: estabilitat i resincronitzacions

Aquesta secció presenta els resultats sobre àudio real de les quatre parelles de fonts heterogènies. Com que els clips no comparteixen el mateix instant d'inici (veure limitació Secció 3.1), no disposem de ground truth. La mètrica clau és l'estabilitat: un sistema que detecta consistentment el mateix retard en totes les finestres és fiable. Es mesuren la desviació

estàndard (std) i les resincronitzacions (rsy): salts > 0,5 s entre finestres consecutives perceptibles a l'àudio de sortida.

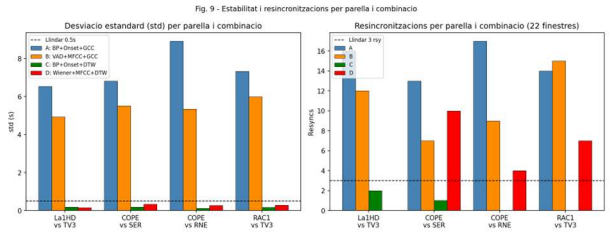


Fig. 9. Estabilitat (std) i resincronitzacions per parella i combinació. A i B (GCC-PHAT): alta variabilitat. C i D (DTW): molt consistents.

TAULA V. ESTABILITAT (std/rsy) PER 22 FINESTRES I RETARD MIG DETECTAT

Parella	A std/rsy	B std/rsy	C std/rsy	D std/rsy	C ret.med (s)	D ret.med (s)
La1HD–TV3	6,54/16	4,93/12	0,19/2	0,16/0	0,45	0,41
COPE–SER	6,82/13	5,51/7	0,18/1	0,33/10	0,40	0,67
COPE–RNE	8,92/17	5,35/9	0,12/0	0,26/4	0,42	0,51
RAC1–TV3	7,33/14	5,99/15	0,17/0	0,28/7	0,51	0,72
Mitjana	7,40/15	5,44/11	0,17/1	0,26/5	0,45	0,58

5.5 Interpretació del retard real estimat

El retard detectat entre La1HD i TV3 (0,45 s per C, 0,41 s per D, columna 'C/D ret.med.' de la Taula V) és el resultat més directament interpretable com a retard tecnològic real: (1) ambdues fonts són de la mateixa tecnologia (satèl·lit/TDT), minimitzant les diferències de textura acústica; (2) l'àudio és la mateixa senyal retransmesa per dos múltiplexors, generant un offset tecnològic pur.

Cal tenir en compte que aquest valor de 0,41–0,45 s és la suma del retard tecnològic pur més qualsevol diferència de punt d'inici de gravació. Per aïllar el retard tecnològic pur caldria un punt de referència temporal comú (treball futur). Per a les parelles heterogènies (ràdio–TV), la seva estabilitat (std < 0,35 s) confirma que el sistema detecta consistentment algun desplaçament real entre les dues fonts.

5.6 Influència del tipus de transmissió en els errors

Els experiments revelen un patró clar: la dificultat de la sincronització augmenta amb la heterogeneïtat tecnològica de la parella. COPE–RNE (dues ràdios en castellà per internet) és la parella més estable per a C (std=0,12 s, 0 resyncs), però la més inestable per a GCC-PHAT (std=8,92 s, 17 resyncs).

Una resincronització (resync) es produeix quan el retard detectat en una finestra difereix en més de 0,5 s del de la finestra anterior, indicant que el sistema ha perdut la pista del retard i ha de recalibrar. La causa paradoxal dels 17 resyncs de GCC-PHAT en COPE–RNE s'explica per la periodicitat del senyal: dues ràdios amb el mateix locutor i ritme de narració similars generen una funció de correlació molt plana, amb múltiples pics de magnitud molt similar separats per la periodicitat del ritme del discurs. DTW, que alinea frame a frame sense dependre de cap pic global, no pateix aquest efecte.

Això té una implicació pràctica: GCC-PHAT seria més adequat per a parelles TV–TV del mateix canal (senyal gairebé idèntic, pic més clar) que no pas per a parelles TV–ràdio o ràdio–ràdio de fonts similars.

5.7 Anàlisi de portabilitat i cost computacional

L'anàlisi s'ha executat sobre un PC de sobretaula amb Windows en un entorn Python 3. S'ha escollit Python perquè les biblioteques librosa i scipy ja venen implementades en aquest llenguatge i, a més, és fàcilment traslladable a plataformes més petites basades en ARM, com la Raspberry Pi 4, que és el target objectiu en un futur. Per executar el pipeline sense modificar el codi cal: (1) sistema operatiu Linux o Windows, (2) MMU per suportar Python, (3) suport per a numpy, scipy i librosa. Això inclou la Raspberry Pi 4 i la Raspberry Pi Zero 2W (ambdues amb Linux/Python disponibles), però exclou el STM32 Cortex-M (sense Linux ni Python; requeriria traducció del codi a C i reimplementació de les biblioteques).

La combinació C és la millor candidata per a portabilitat: 4.122 crides cProfile vs 35.401 de D, i no depèn del filtre Wiener (coll d'ampolla de D amb 58% del temps). Com a treball futur, executar el profiling directament en una Raspberry Pi via SSH permetria validar la portabilitat amb dades reals.

5.8 Resum comparatiu de totes les mètriques

TAULA VI. RESUM COMPARATIU DE TOTES LES MÈTRIQUES

Criteri	A: Onset+ GCC	B: MFCC+ GCC	C: Onset+ DTW	D: MFCC+ DTW
Error GT mig	5.451 ms	3.498 ms	traceback	traceback
Estabilitat real (std)	7,40 s	5,44 s	0,17 s	0,26 s
Resyncs mig / 22	15	11	1	5
T/finestra (ms)	19,8	28,7	22,8	60,6
Crides cProfile	2.497	13.301	4.122	35.401

El problema del traceback DTW apareix quan el retard és gran: el camí òptim tendeix a la diagonal zero i retorna valors propers a 0 s en lloc del retard real. La solució (treball futur) és la banda de Sakoe-Chiba.

6 Discussió

La divergència entre resultats sintètics i reals és la principal conclusió metodològica: GCC-PHAT millor en laboratori, pitjor en real; DTW al contrari. Això significa que avaluar un sistema de sincronització només amb ground truth artificial pot dur a conclusions radicalment incorrectes sobre quin algorisme és millor en la pràctica.

La causa d'aquesta inestabilitat és la heterogeneïtat acústica entre les fonts: codificadors, locutors i entorns sonors completament diferents entre ràdio i televisió generen una funció de correlació amb múltiples màxims locals de magnitud similar. Per il·lustrar-ho, la Fig. 5 mostra com en condicions sintètiques GCC-PHAT obté un pic únic i clar, mentre que en àudio real (La1HD vs TV3) apareixen múltiples pics similars que fan impossible determinar quin és el correcte. Fins i tot les parelles homogènies (ràdio–ràdio) pateixen el problema de periodicitat del discurs: dos locutors amb el mateix ritme

narratiu generen una funció de correlació molt plana. DTW no cerca cap pic global sinó que alinea frame a frame minimitzant el cost acumulat, cosa que el fa robust a totes dues dificultats.

El problema del traceback DTW és tècnic i té solució: la restricció de la banda de Sakoe-Chiba [5] limita el camí de traceback a una diagonal de màxima desviació r, forçant l'algorisme a explorar només els retards plausibles i evitant que el camí tendeixi a la diagonal zero. Alternativament, el retard pot extreure's com la posició del punt d'inici del traceback respecte a la diagonal principal.

Substituir el Wiener de D per un Bandpass simple reduiria el temps de processament de forma significativa (eliminant el 58% del temps total), resolent la inviabilitat sense cap pèrdua de robustesa en la majoria de casos, ja que la font d'estabilitat de D és el DTW, no el Wiener. No obstant, aquesta substitució té un pro i un contra: PRO: reducció dràstica del cost computacional i viabilitat en Raspberry Pi. CONTRA: el Wiener elimina el soroll de fons de forma molt més precisa que el Bandpass, especialment útil per a fonts de televisió amb ambient d'estadi. Per tant, la substitució és clarament recomanable per a parelles ràdio–ràdio, però per a parelles que incloguin TV caldria avaluar si la pèrdua de precisió en el filtratge afecta l'estabilitat.

Una std de 0,17 s significa que, en el 68% dels casos (interval d'una desviació estàndard), l'estimació del retard varia menys de 0,17 s d'una finestra a la següent, cosa que és imperceptible a l'oïda. Les resincronitzacions (salts > 0,5 s) són molt més crítiques perquè es perceben com a salts sobtats a l'àudio de sortida. La combinació C minimitza les resincronitzacions a 1 de mitjana sobre les 22 finestres analitzades (120 s d'àudio dividits en finestres de 10 s amb solapament del 50%, tal com s'il·lustra a la Fig. 1).

7 Conclusions

Aquest treball ha demostrat la viabilitat de la sincronització automàtica d'àudio entre fonts esportives heterogènies sense intervenció manual. Les conclusions principals són:

Bandpass+Onset+DTW (C) és el candidat principal: std=0,17 s i una resincronització de mitjana en quatre parelles reals, 22,8 ms/finestra. La portabilitat a Raspberry Pi és viable per la naturalesa Python del codi; la validació real en hardware queda com a treball futur.

GCC-PHAT és inestable en àudio real heterogeni (std 5–9 s) per l'ambigüitat de pic causada per la heterogeneïtat de fonts. Fins i tot parelles homogènies pateixen el problema de periodicitat del discurs. En teoria, aquesta inestabilitat es podria intentar corregir aplicant una restricció de banda de Sakoe-Chiba al traceback del DTW, que limitaria el camí a una franja al voltant de la diagonal i evitaria múltiples pics espuris. Aquesta millora queda com a treball futur.

El profiling real amb cProfile identifica el Wiener com a coll d'ampolla (58% del temps de D) i mostra una millora concreta i directament implementable: substituir Wiener per Bandpass en D reduiria el temps de forma significativa.

L'avaluació únicament en condicions sintètiques és insuficient: GCC-PHAT és el millor en laboratori i el pitjor en la pràctica. Qualsevol sistema d'aquest tipus ha d'avaluar-se obligatòriament sobre àudio real heterogeni.

Agraïments

Vull agrair personalment al meu tutor, en Víctor Montilla, ja que sense el seu suport en cada reunió aquest projecte no hagués vist la llum. La seva dedicació i ajuda ha estat crucial per poder aconseguir l'objectiu proposat.

Bibliografia

[1] AWS (s.d.). Latencia de vídeos en streaming en directo. Amazon Web Services. https://aws.amazon.com/es/media/tech/video-latency-in-live-streaming/

[2] Knapp, C.H. & Carter, G.C. (1976). The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech and Signal Processing, 24(4), 320–327.

[3] Wang, A. (2003). An industrial-strength audio search algorithm. Proceedings of the 4th ISMIR Conference, Baltimore.

[4] Davis, S. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition. IEEE TASLP, 28(4), 357–366.

[5] Sakoe, H. & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. IEEE TASLP, 26(1), 43–49.

[6] Bello, J.P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M. & Sandler, M.B. (2005). A tutorial on onset detection in music signals. IEEE TASLP, 13(5), 1035–1047.

[7] Ephraim, Y. & Malah, D. (1984). Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE TASLP, 32(6), 1109–1121.

[8] McFee, B., Raffel, C., Liang, D. et al. (2015). librosa: Audio and music signal analysis in Python. SciPy Conference, Austin.

[9] Salamon, J. & Bello, J.P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE SPL, 24(3), 279–283.

[10] Sohn, J., Kim, N.S. & Sung, W. (1999). A statistical model-based voice activity detection. IEEE SPL, 6(1), 1–3.

[11] FEED Magazine (2025). Losing the lag: Why sports streaming is one of the biggest victims of delay. https://feedmagazine.tv/signal/latency-in-sports-losing-the-lag/

APÈNDIX

A. Arxius d'àudio originals i clips processats

La Taula A-I llista els set arxius originals enregistrats simultàniament durant la final de la Copa del Rei 2026, juntament amb els clips WAV extrets per als experiments. Tots els clips són monocanal, 16.000 Hz, 16 bits, 120 s de duració.

TAULA A-I. ARXIUS ORIGINALS I CLIPS WAV PROCESSATS

Arxiu original	Font	For mat	Mida orig.	Clip WAV	Mida WAV
20260418_La 1HD.ts	La 1 HD	.ts	386 MB	La1HD_2m in.wav	3,66 MB
20260418_La 1UHD.ts	La 1 UHD	.ts	1,4 GB	(no usat)	—
20260418_TV 3.ts	TV3	.ts	639 MB	TV3_2min. wav	3,66 MB
18042026_CO PE.mp4	COPE	.mp 4	57 MB	COPE_2mi n.wav	3,66 MB
18042026_RA C1.mp4	RAC1	.mp 4	56 MB	RAC1_2mi n.wav	3,66 MB
18042026_RN E.mp4	RNE	.mp 4	56 MB	RNE_2min. wav	3,66 MB
18042026_SE R.mp4	SER	.mp 4	56 MB	SER_2min. wav	3,66 MB

La 1 UHD no s'ha inclòs als experiments per la seva resolució UHD (arxiu massa gran); el còdec d'àudio és idèntic a La1HD.

B. Resultats complets d'estabilitat per parella i combinació

La Fig. A-1 mostra de forma ampliada els resultats d'estabilitat i resincronitzacions per a totes les parelles i combinacions. El llindar de la línia discontinua representa el límit acceptable: std=0,5 s i 3 resincronitzacions.

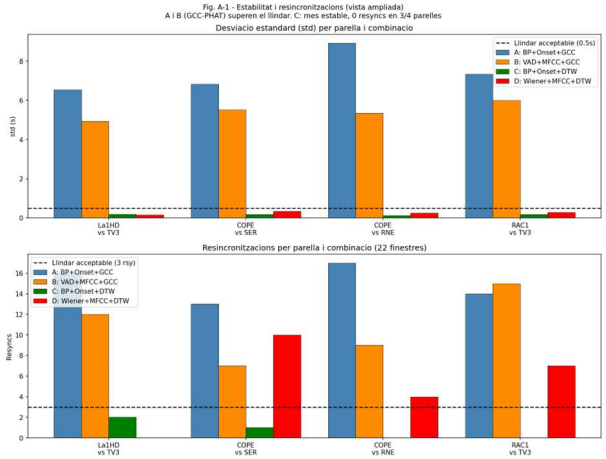


Fig. A-1. Estabilitat i resincronitzacions per a les quatre parelles reals i les quatre combinacions. Les combinacions A i B (GCC-PHAT) superen amb escreix el llindar acceptable. C és la més estable i la única amb 0 resyncs en 3 de 4 parelles.

C. Paràmetres del pipeline i justificació detallada

La Taula A-II detalla tots els paràmetres del pipeline, els valors seleccionats i la justificació de cada decisió de disseny. Els paràmetres marcats amb (*) s'han obtingut experimentalment (grid search o averaging).

TAULA A-II. PARÀMETRES DEL PIPELINE, VALORS I JUSTIFICACIÓ. (*) = OBTINGUT EXPERIMENTALMENT

Paràmetre	Valor	Justificació
Freq. mostreig (SR)	16.000 Hz	Cobreix fins a 8 kHz per Nyquist. Suficient per a veu i sons esportius. Meitat del cost de 32 kHz.
Mida finestra W (*)	10 s	Grid search sobre [3, 5, 7, 10, 12, 15 s] minimitzant error GT=5 s. Millor compromís precisió/latència.
Pas finestra (*)	5 s (50%)	50% solapament garanteix cobertura per ≥2 finestres consecutives. Latència màx: 5 s + temps procés.
Hop size (Onset/MFCC)	512 mostres (32 ms)	Estàndard en librosa [8]. Balanç entre resolució temporal i cost computacional.
Freqs. Bandpass	100–7.900 Hz	Límit inferior: elimina ronc elèctric (50 Hz). Límit superior: soroll d'alta freq. irrellevant per a veu.
Ordre Butterworth	4	Atenuació -24 dB/octava. Suficient sense artefactes de fase significatius.
Coeficients MFCC	13	Estàndard en reconeixement de parla [4]. Més coeficients no milloren la comparació entre fonts.
Max. frames DTW	300	Limita el cost $O(N^2)$ a 300×300=90.000 operacions. Equivalent als primers ~9,6 s d'Onset.
Averaging N (*)	5	N=5 redueix variabilitat per inicialització de biblioteques sense augmentar la latència real.
Llindar resync	>0,5 s	Salt perceptible a l'àudio de sortida. Canvis inferiors a 0,5 s són imperceptibles a l'oïda.

D. Millores futures identificades

La Taula A-III llista les millores futures, ordenades de menys a més complexitat d'implementació.

TAULA A-III. MILLORES FUTURES IDENTIFICADES, ORDENADES PER COMPLEXITAT

Millora	Descripció tècnica	Impacte esperat	Complexitat
Substitució Wiener→Band pass en D	Eliminar filtre Wiener de D i substituir pel Bandpass simple.	D viable en RPi. Sense pèrdua de robustesa (l'estabilitat ve del DTW).	Molt baixa
Correcció traceback DTW (Sakoe-Chiba)	Implementar banda r al traceback: el camí no pot allunyar-se >r frames de la diagonal.	Correcció de l'error estructural en GT. Valida C també en condicions sintètiques.	Baixa
Talls TV pel rellotge de pantalla	Usar el rellotge de transmissió visible a la imatge TV per sincronitzar els talls.	Permet calcular el retard tecnològic pur TV–TV eliminant l'offset d'enregistrament.	Baixa
Validació RPi 4 físic amb cProfile	Executar el pipeline en RPi 4 via SSH i mesurar temps i memòria reals.	Substitueix els factors teòrics per mesures reals. Necessari per a un producte.	Mitjana
Audio Fingerprinting (Shazam [3])	Implementar extracció d'empremta digital basada en constel·lació de pics espectrals.	Més robust per a parelles molt heterogènies. Possible millora sobre GCC-PHAT.	Alta
CNN/LSTM per reconeixement de transitoris [9]	Entrenar una xarxa neuronal per detectar transitoris comuns invariants a les diferències de font.	Màxima robustesa en entorns molt sorollosos. Aprèn invariàncies del context esportiu.	Molt alta