

TEXNO: extracción heurística del ADN estructural de composiciones musicales

Pipeline DSP para análisis de forma, tempo, espectro, armonía y dinámica

Motor TEXNO v1.1 · Documentación técnica
Gustavo Arteaga · Proyecto open source · Julio 2026
<https://github.com/GustavoUNAL/TEXNO>

RESUMEN

TEXNO es un sistema de análisis musical que convierte señales de audio en una representación estructurada serializable (JSON/YAML) denominada *ADN de la pista*. El motor implementa un pipeline de procesamiento digital de señales (DSP) sin aprendizaje automático: decodificación PCM, mezcla mono, envolvente de energía RMS, autocorrelación temporal para tempo, transformada de Fourier de tiempo discreto (DFT/FFT) con ventana de Hann, agregación espectral logarítmica, cromagrama de 12 clases de altura, estimación de tonalidad por correlación con perfiles Krumhansl-Schmuckler, segmentación de forma por energía relativa y heurísticas de capas instrumentales. Este documento describe con detalle los algoritmos, parámetros, limitaciones y el esquema de salida. El sistema está orientado a música electrónica y dance, donde la repetición rítmica y el espectro comprimido facilitan estimaciones robustas, aunque el marco es aplicable a otros géneros con las reservas indicadas.

Palabras clave:

MIR, análisis musical, FFT, cromagrama, estimación de tonalidad, BPM, segmentación de forma, techno.

1. Introducción

El análisis musicológico y la ingeniería de audio tradicionalmente requieren herramientas especializadas (DAW, espectroanalizadores, detectores de tempo). TEXNO unifica en un único flujo automático la extracción de descriptores de bajo y medio nivel que caracterizan una grabación masterizada: tempo, distribución espectral, centroide, dinámica, estructura macro (intro, build, drop, break, outro), pitch classes dominantes y capas estimadas (kick, bajo, percusión, hats, aire).

La salida no pretende reemplazar la escucha crítica ni la separación de fuentes por modelos neuronales; constituye una *firma estructural* reproducible, útil para catalogación, comparación de mixes, documentación de sets y prototipado de herramientas de remix o recomendación.

2. Arquitectura del sistema

El pipeline se organiza en cinco etapas secuenciales:

1. **Ingesta y metadatos** — decodificación con `@audio/decode`, lectura ID3/container con `music-metadata`.
2. **Preprocesamiento** — mezcla a mono, normalización implícita por RMS frame.

3. **Análisis temporal** — envolvente, BPM, onsets, curva de energía.
4. **Análisis espectral y armónico** — FFT, bandas, curva log-frecuencia, cromagrama, tonalidad.
5. **Inferencia estructural** — forma, capas, narrativa, serialización ADN.

```
Audio → decode → mono[n] → envelope[] → {BPM, onsets, form, energy_curve}
      ↘ FFT blocks → {spectrum, chroma, harmony, layers}
```

3. Representación de la señal

3.1 Decodificación y mezcla mono

Sea $x_c[n]$ la señal discreta del canal c a frecuencia de muestreo f_s . La mezcla mono se define como:

$$y[n] = (1/C) \cdot \sum_{c=1}^C x_c[n]$$

donde C es el número de canales. Esta reducción elimina información de panorámica pero estabiliza estimadores espectrales y de energía en material estéreo ampliamente correlacionado (típico en masters de club).

3.2 Envolvente de energía (RMS por tramas)

La señal se divide en tramas de $N_f = \lfloor f_s \cdot T_f \rfloor$ muestras con $T_f = 50 \text{ ms}$ y salto igual al tamaño de trama (sin solapamiento). Para cada trama k :

$$E[k] = \sqrt{(1/N_f) \cdot \sum_{i=0}^{N_f-1} y[n_k+i]^2}$$

La secuencia $\{E[k]\}$ constituye la envolvente de amplitud y es la base para tempo, onsets, segmentación de forma y curva de energía narrativa.

4. Análisis temporal y rítmico

4.1 Estimación de tempo (BPM)

Se aplica autocorrelación no normalizada sobre la envolvente centrada en media: $\tilde{E}[k] = E[k] - \mu_E$. Para cada retardo τ en el rango compatible con 70–180 BPM:

$$R(\tau) = \sum_k \tilde{E}[k] \cdot \tilde{E}[k+\tau]$$

El retardo τ^* que maximiza $R(\tau)$ se convierte en tempo $BPM = 60 \cdot f_{env} / \tau^*$, donde $f_{env} = 1000/T_f$ es la tasa de frames de la envolvente. Se aplican reglas de octava para favorecer el rango electrónico (duplicar si BPM < 100, dividir si BPM > 160). La confianza se escala linealmente con $R(\tau^*)$ respecto a la longitud de la envolvente.

4.2 Detección de onsets

Picos locales en $E[k]$ por encima de $\theta = 1.4 \cdot \mu_E$, con separación mínima de 80 ms entre eventos. La densidad rítmica se reporta como *onsets / duración* (eventos por segundo).

5. Análisis espectral

5.1 Bloques FFT

La señal se analiza en bloques de $N = 2048$ muestras con ventana de Hann:

$$w[n] = 0.5 \cdot (1 - \cos(2\pi n / (N-1)))$$

El avance entre bloques es $hop = 4N$ muestras (~ 186 ms a 44.1 kHz), reduciendo coste computacional. Se implementa FFT radix-2 in-place (Cooley–Tukey) sobre las componentes real e imaginaria.

Para cada bin k , la magnitud es $|X[k]| = \sqrt{Re^2 + Im^2}$ y la frecuencia asociada $f_k = k \cdot f_s / N$.

5.2 Bandas de frecuencia

La energía espectral se agrega en cinco bandas estándar de mezcla:

Banda	Rango (Hz)	Interpretación
SUB	20–60	Sub-graves, kick fundamental
LOW	60–250	Bajo, cuerpo
MID	250–2000	Percusión, snare, cuerpo de sintes
HIGH	2000–8000	Presencia, hi-hats
AIR	8000–20000	Brillo, ruido, FX

Cada banda se normaliza dividiendo por la suma total de magnitudes en el rango audible (20 Hz–20 kHz), obteniendo proporciones en $[0,1]$.

5.3 Curva espectral logarítmica

Para visualización de espectro completo se definen $B = 96$ bins logarítmicos entre 20 Hz y 20 kHz. Los bordes de bin siguen:

$$f_i = f_{min} \cdot (f_{max}/f_{min})^{i/B}$$

La magnitud por bin se acumula y normaliza por el máximo global, reportándose también en dB relativos: $20 \log_{10}(mag)$.

5.4 Centroide espectral

$$SC = \sum_k f_k \cdot |X[k]| / \sum_k |X[k]|$$

Valores altos del centroide (> 4 kHz) se clasifican como brillo *high*; entre 2–4 kHz *medium*; por debajo *low*.

6. Análisis armónico

6.1 Cromagrama

Para frecuencias entre 55 Hz y 5 kHz, cada bin espectral se proyecta a una clase de altura (pitch class) de 12 semitonos mediante:

$$MIDI(f) = 69 + 12 \cdot \log_2(f / 440)$$

La contribución al cromagrama se pondera por magnitud, con factor 1.35 para $f < 400$ Hz (énfasis en fundamentales) y 0.7 para $f > 2$ kHz (atenuación de parciales agudos). El vector de 12 dimensiones se normaliza a suma unitaria.

6.2 Estimación de tonalidad (Krumhansl–Schmuckler)

Se correlaciona el cromagrama con perfiles de sensación de tonalidad para modo mayor y menor (Krumhansl & Kessler, 1982; implementación con correlación de Pearson circular para las 12 transposiciones). La clave y modo con mayor correlación definen la tonalidad estimada. La confianza se mapea desde la correlación máxima:

$$conf = clamp((r_{max} - 0.55) / 0.4, 0, 1)$$

En material electrónico monotonal o altamente percusivo, la confianza armónica puede ser alta estadísticamente pero musicalmente ambigua; debe interpretarse junto al espectro y al género.

6.3 Picos espectrales y notas

Por bloque FFT se registra el pico de mayor magnitud entre 40 Hz y 2 kHz. Los picos globales más fuertes se mapean a notas (nombre + octava) evitando duplicar pitch classes, hasta 8 candidatos, con fuerza relativa normalizada.

7. Dinámica

Sobre la señal mono completa:

- **Peak** — máximo $|y[n]|$ en escala lineal y dBFS: $20 \log_{10}(\text{peak})$.
- **RMS** — $\sqrt{(\sum y[n]^2 / N)}$.
- **Rango dinámico** — $\text{peak}_{\text{dB}} - \text{RMS}_{\text{dB}}$.
- **LUFS aproximado** — $\text{RMS}_{\text{dB}} - 0.691$ (heurística, no EBU R128).

La *roughness* timbral se categoriza por rango dinámico: < 6 dB *compressed*, 6–12 dB *moderate*, > 12 dB *open*.

8. Segmentación de forma musical

La envolvente se agrupa en ventanas de $\approx 1/16$ de la duración total. Cada ventana recibe energía normalizada $\hat{e} = e / \max(e)$ y una etiqueta heurística:

Condición	Etiqueta
$t < 12\%$ duración y $\hat{e} < 0.45$	intro
$t > 85\%$ duración	outro
$\hat{e} < 0.35$	break
$\hat{e} < 0.55$	build
en otro caso	drop

Secciones contiguas con la misma etiqueta se fusionan. El número de compases se estima dividiendo la duración de la sección por $4 \cdot 60/\text{BPM}$ (compás 4/4 asumido). La energía por sección se reporta en escala 1–10.

9. Estimación de capas (heurística)

A partir de las proporciones espectrales y el BPM, se infiere presencia de capas instrumentales sin separación de fuentes:

- $\text{SUB} > 8\% \rightarrow \text{kick}$
- $\text{LOW} > 10\% \rightarrow \text{bass}$
- $\text{MID} > 15\% \rightarrow \text{percussion}$
- $\text{HIGH} > 12\% \rightarrow \text{hats}$
- $\text{AIR} > 8\% \rightarrow \text{air_fx}$
- $\text{BPM } 130\text{--}160 \rightarrow \text{synth_texture}$ (contexto techno)

La presencia se clasifica como *high*, *medium* o *low* según umbrales relativos dentro de cada banda.

10. Esquema de salida (ADN)

El objeto JSON/YAML *Track DNA* contiene los siguientes módulos:

Módulo	Campos principales
track	title, bpm, duration_s, key, sample_rate, codec
form	array de {name, start_s, end_s, bars, energy}
mix	bands, spectrum_curve[], spectral_centroid_hz, rms_db, peak_db
harmony	key, scale, chromagram[12], peaks[], explanation_lines[]
rhythm	onset_times_s[], density_events_per_s, bpm_confidence
timbre	brightness, roughness
narrative	energy_curve[], atmosphere, dominant_element
layers	id, role, freq_hint, presence, notes
analysis_meta	engine, version, analyzed_at, confidence

11. Complejidad y parámetros

Parámetro	Valor
Tamaño FFT N	2048
Hop espectral	$4N$ muestras
Trama RMS	50 ms
Bins curva log	96
Rango BPM	70–180 (con snapping de octava)
Umbral onset	$1.4 \times$ media de envolvente

Para una pista de duración D segundos a 44.1 kHz, el número de bloques FFT es $O(D \cdot f_s / (4N)) \approx O(5.4 \cdot D)$ bloques por minuto de audio. La FFT domina el coste por bloque con $O(N \log N)$.

12. Limitaciones y trabajo futuro

- No hay separación de fuentes (stems); las capas son inferencias espectrales.
- La tonalidad falla en audio atonal, ruido puro o mezclas muy comprimidas.
- El tempo por autocorrelación de envolvente confunde mitades/dobles del pulso real.
- La forma musical es heurística; no detecta letras ni cambios armónicos finos.
- LUFS reportado no cumple norma ITU-R BS.1770.

Extensiones previstas: modelos de separación (Demucs, Spleeter), detección de acordes por HMM/CNN, segmentación con auto-similitud (checkerboard kernel), y calibración perceptual EBU R128.

13. Referencias

1. Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, 89(4), 334–368.
2. Cooley, J. W., & Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90), 297–301.
3. Bello, J. P., et al. (2005). A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5), 1035–1047.
4. Müller, M. (2015). *Fundamentals of Music Processing*. Springer.
5. ITU-R BS.1770 — Algorithms to measure audio programme loudness.