Naturalidad estimada y prosodia

Chatterbox Multilingual · FP32 / Q8 / Q6 · 9 octubre 2026

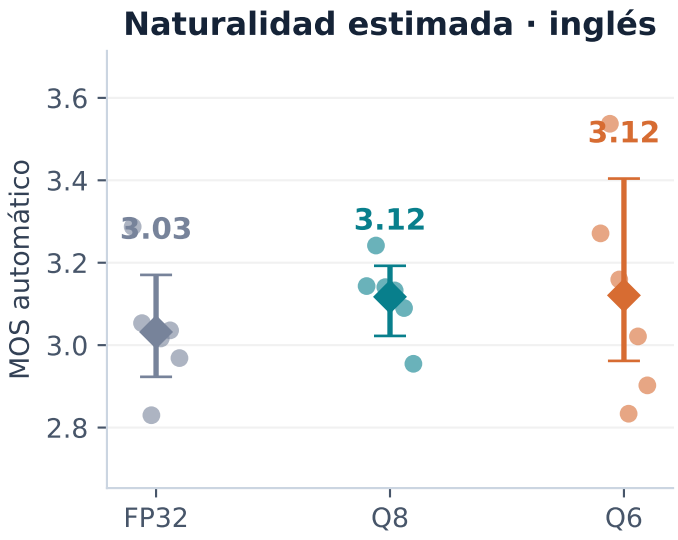
36 clips existentes · 6 frases × 2 semillas · misma voz de referencia

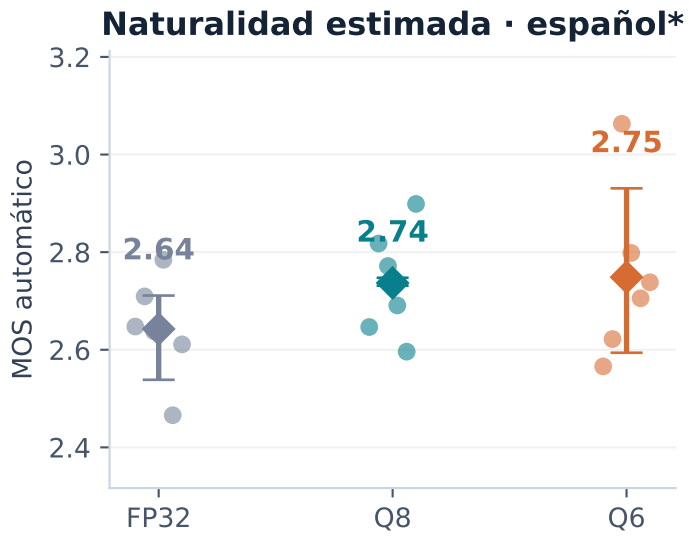
Medición	FP32	Q8	Q6
Naturalidad estimada · inglés	3.032	3.117	3.121
Naturalidad estimada · español (exploratoria)	2.643	2.737	2.749
Variación tonal · DE en semitonos	2.617	2.567	2.632
Rango tonal · P95–P5 en semitonos	8.874	8.552	8.890
Pitch mediano · Hz	215.308	219.872	219.545
HNR medio · dB	14.779	14.850	14.856
Palabras del texto/min	215.792	210.035	209.511
Pausas internas ≥120 ms · %	0.963	1.780	1.460
Rango de intensidad · dB	25.614	25.151	26.224

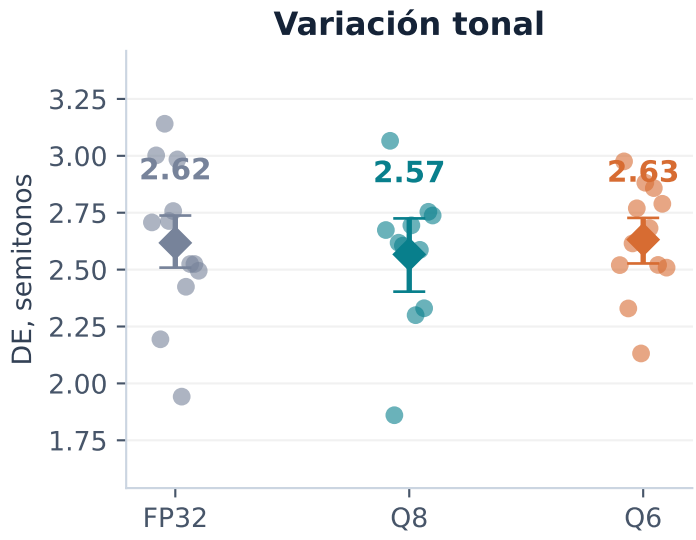
UTMOSv2: 5 folds × 5 crops por clip; MOS automático, sin oyentes.
Español: exploratorio. Solo tres frases por idioma.
Más variación tonal no implica más naturalidad.
CSV, JSON, métodos y fuentes: REPORT.md / REPORT.html.

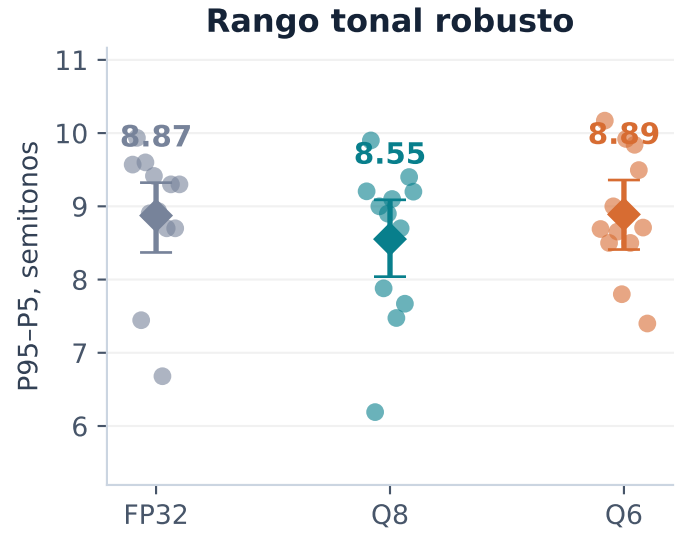
Chatterbox: naturalidad estimada y prosodia

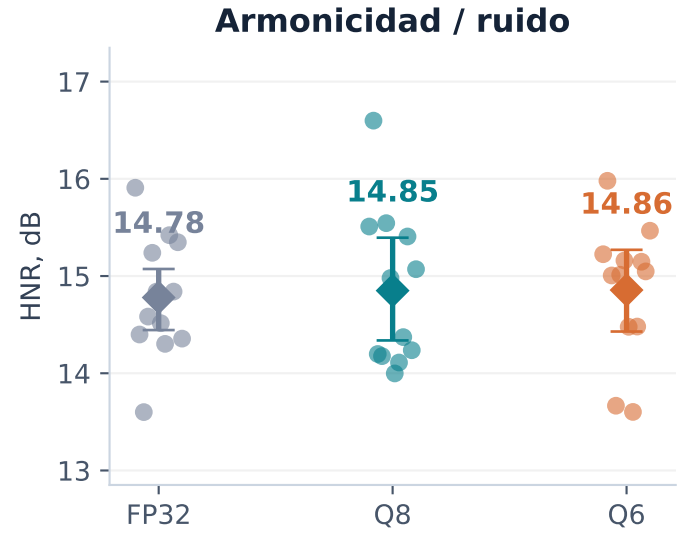
FP32, Q8 y Q6 · 36 clips · 6 frases × 2 semillas · misma referencia de voz

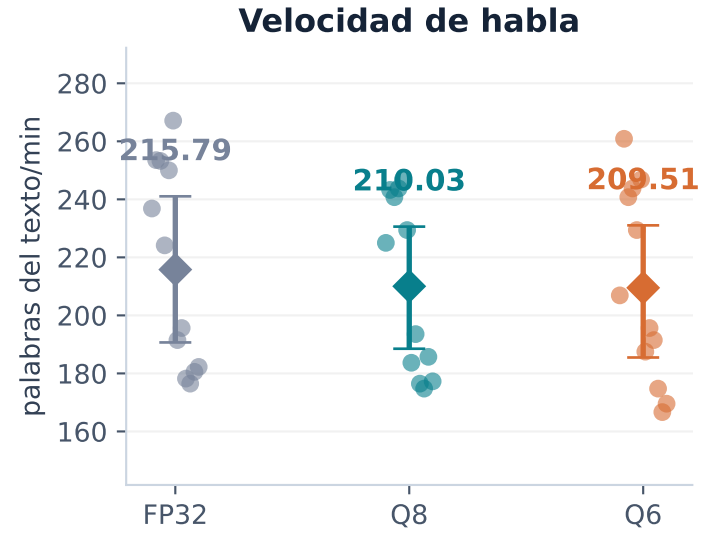








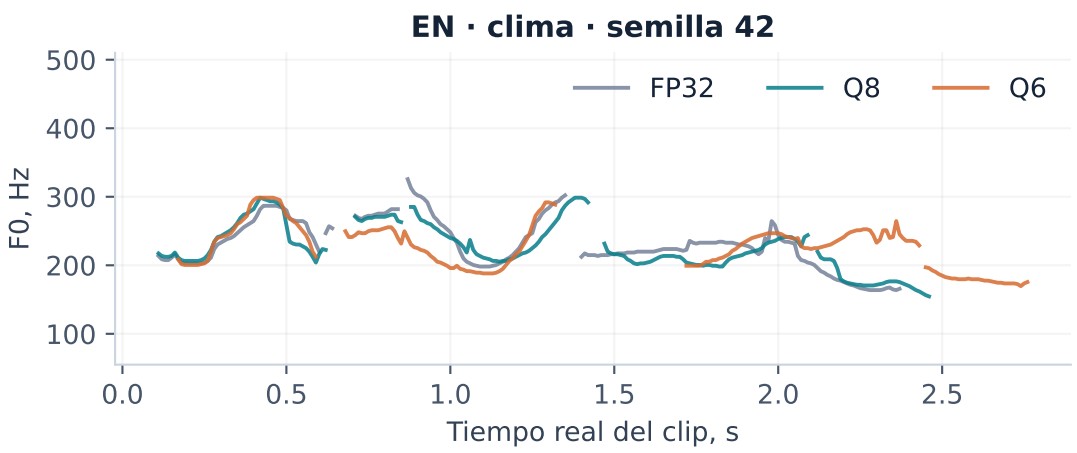


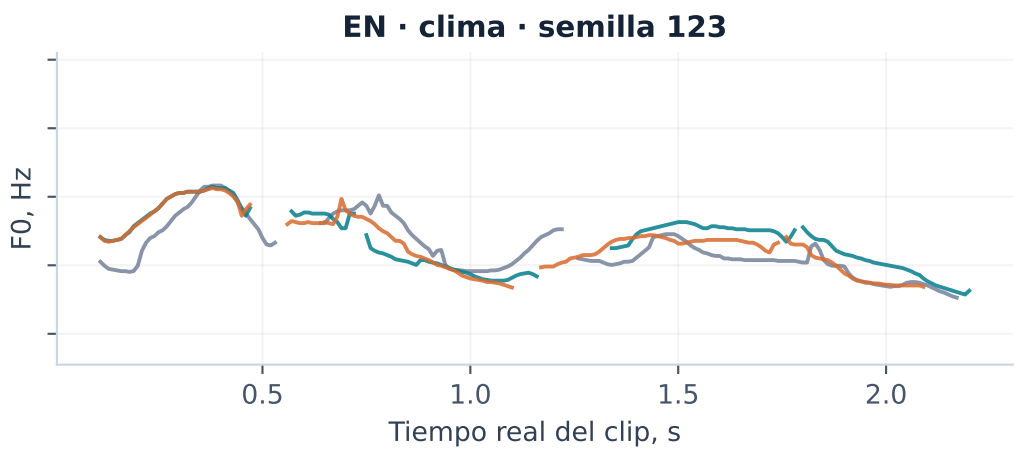


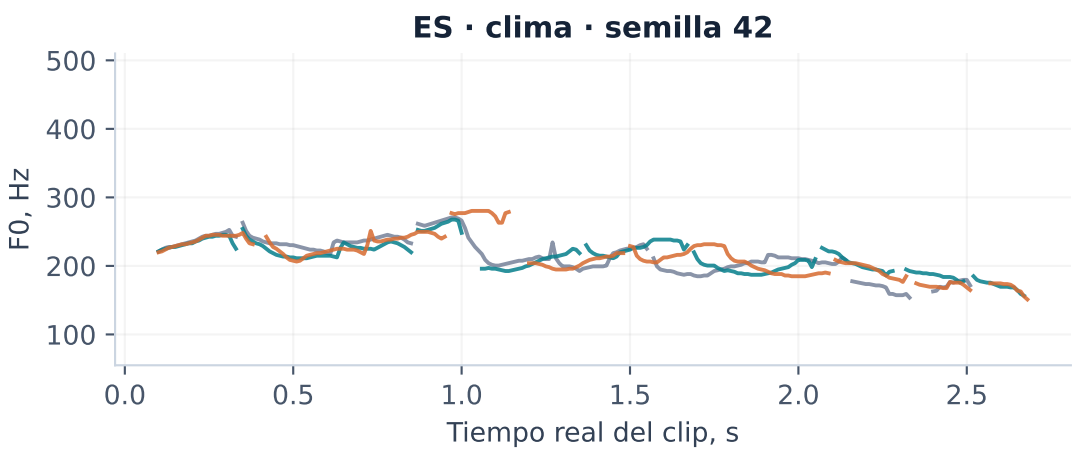
Puntos: clips. Diamantes: media. Barras: bootstrap por frase (95%, solo 3–6 frases).
*Español: puntuación exploratoria. UTMOSv2 no es un MOS humano. Más variación tonal no implica más naturalidad.

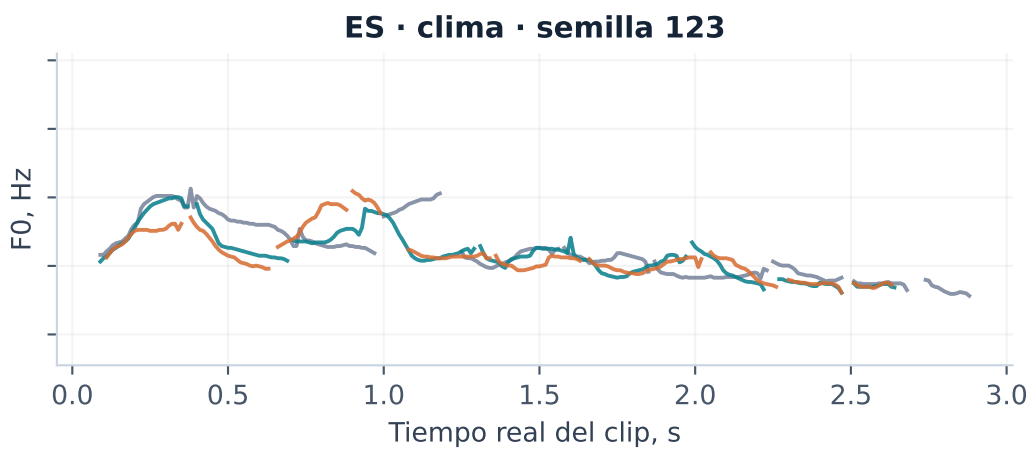
Contornos de pitch · todas las frases y semillas

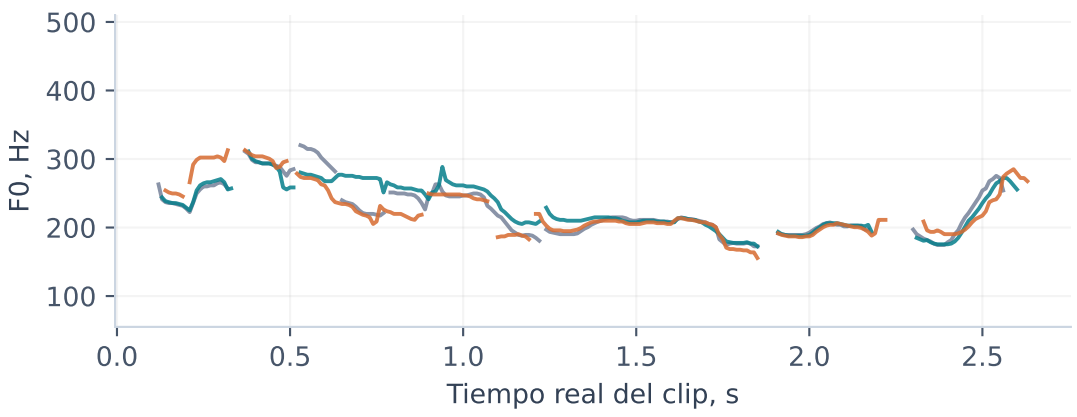
pYIN · 65–500 Hz · paso 10 ms · huecos = sin pitch o energía insuficiente



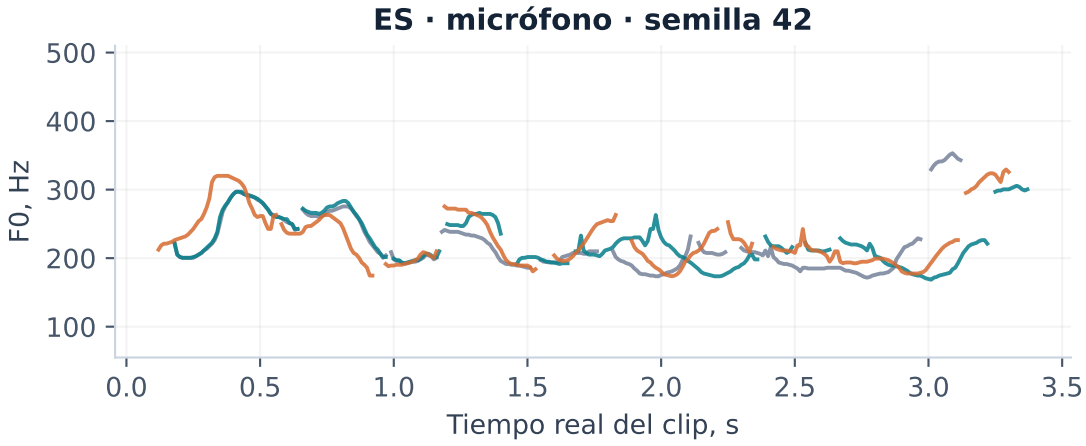


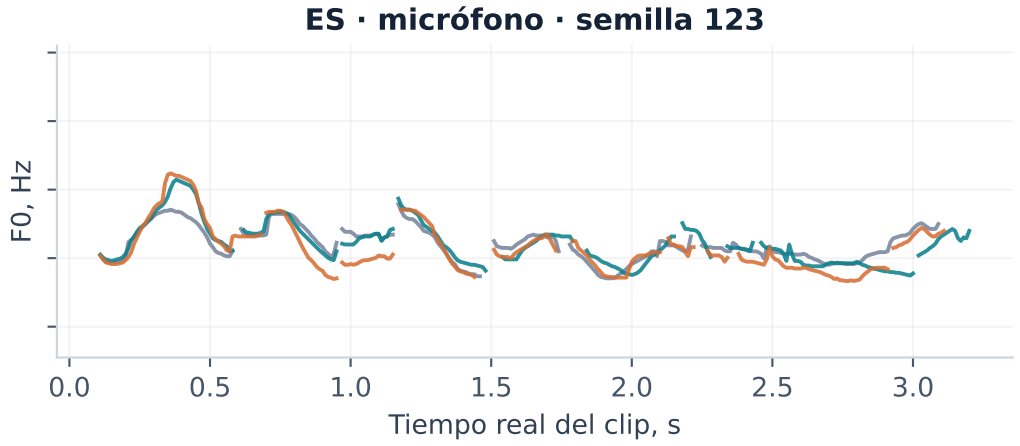


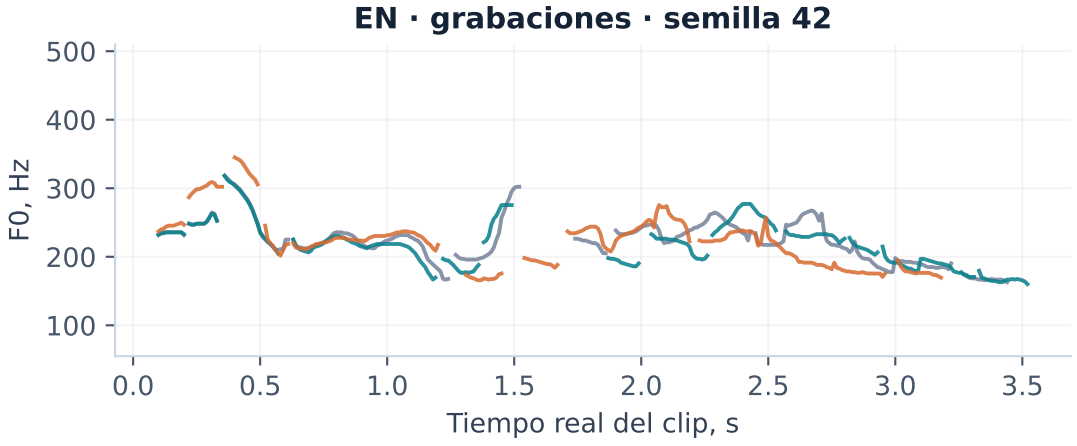


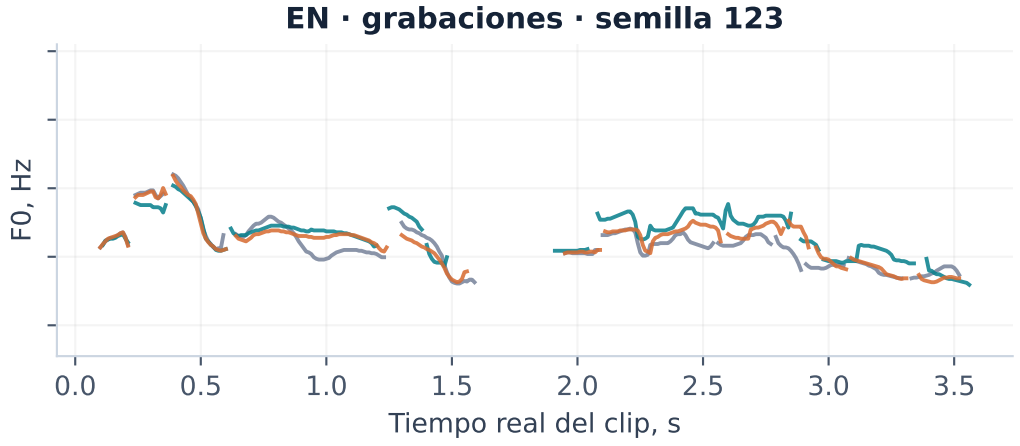


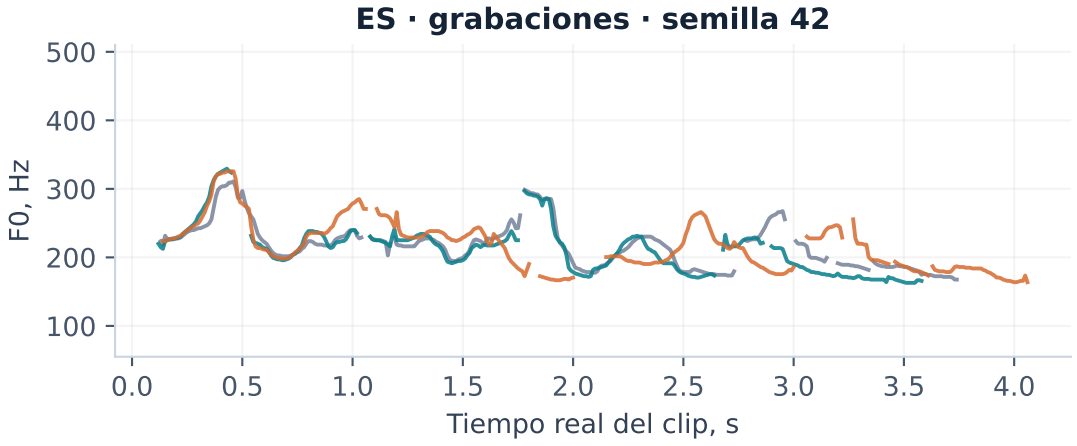


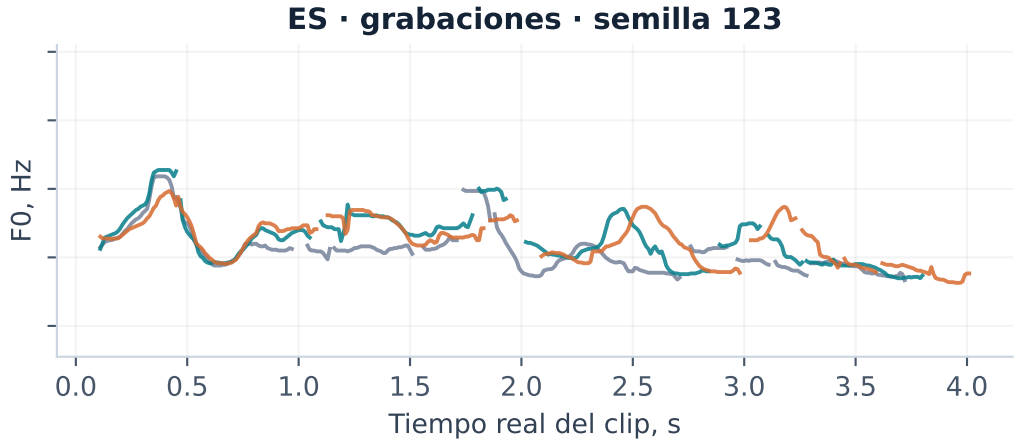








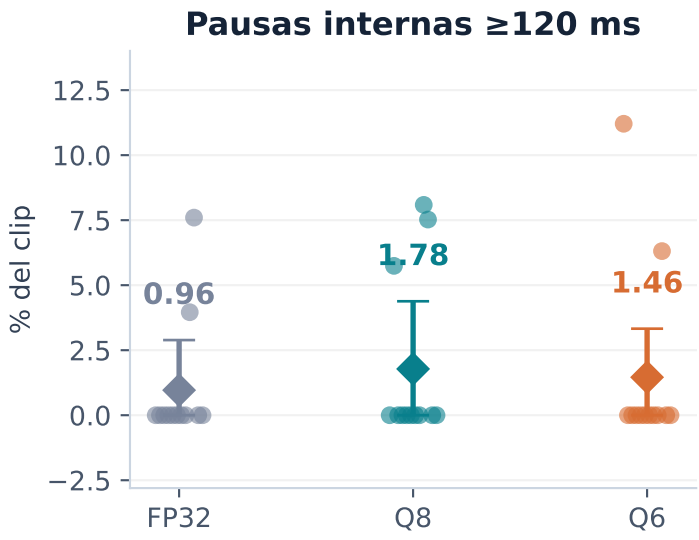


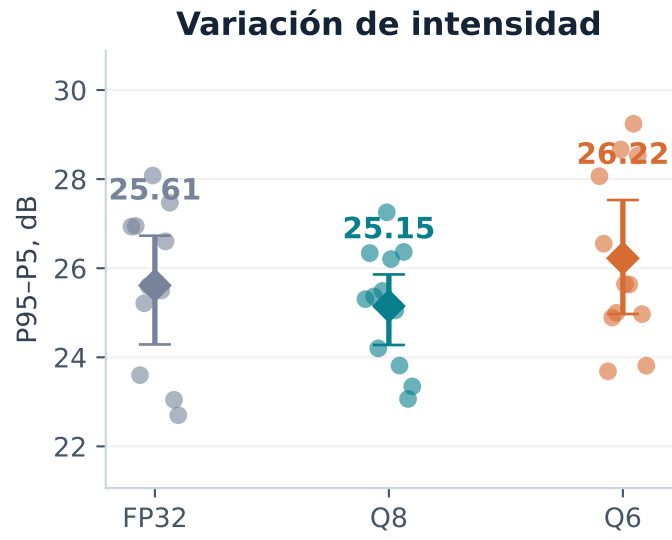


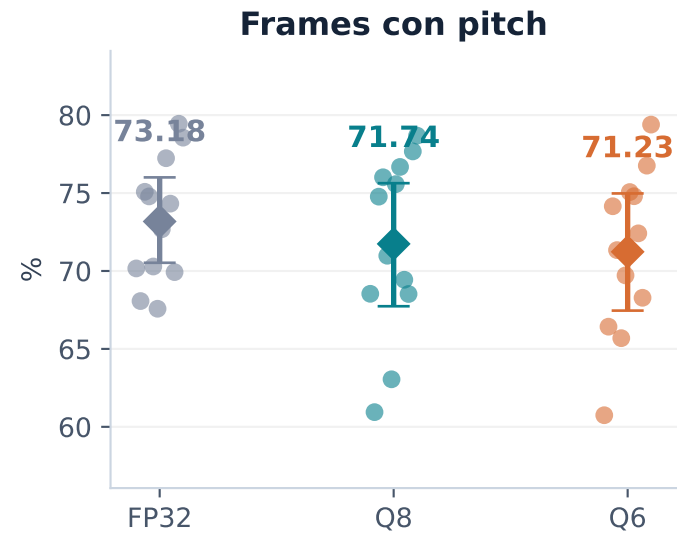
El eje temporal no está alineado por palabras: las diferencias incluyen cambios de duración, pausas y contenido.
Las caídas o saltos pueden reflejar voz crepitante o errores de seguimiento; no se asigna una nota de naturalidad a partir de ellos.

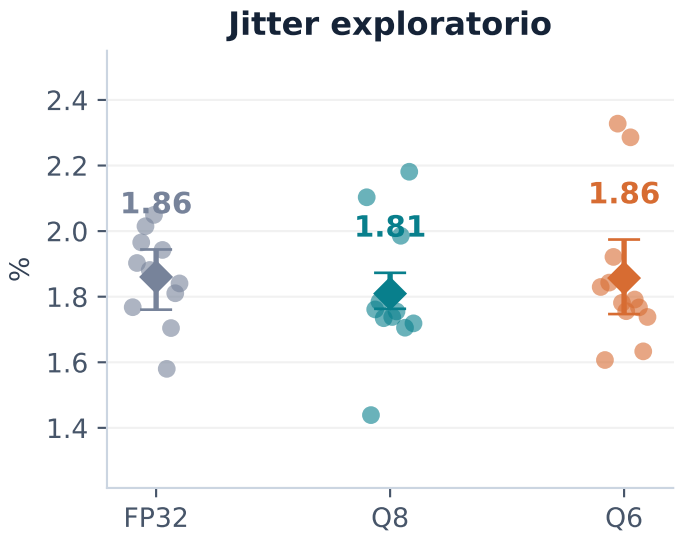
Ritmo, intensidad y estabilidad acústica

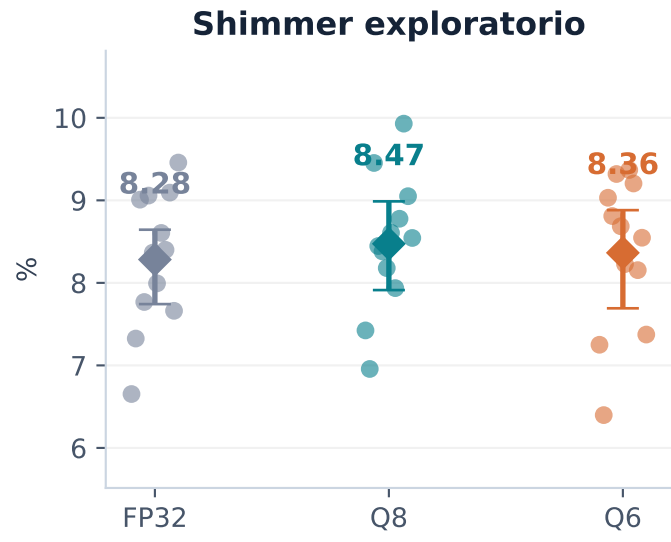
Mismos 36 clips · pausas definidas por energía, no por alineación lingüística

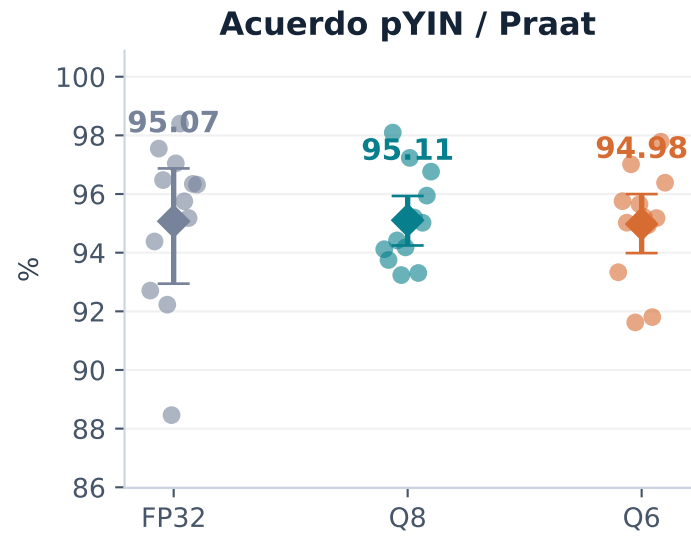








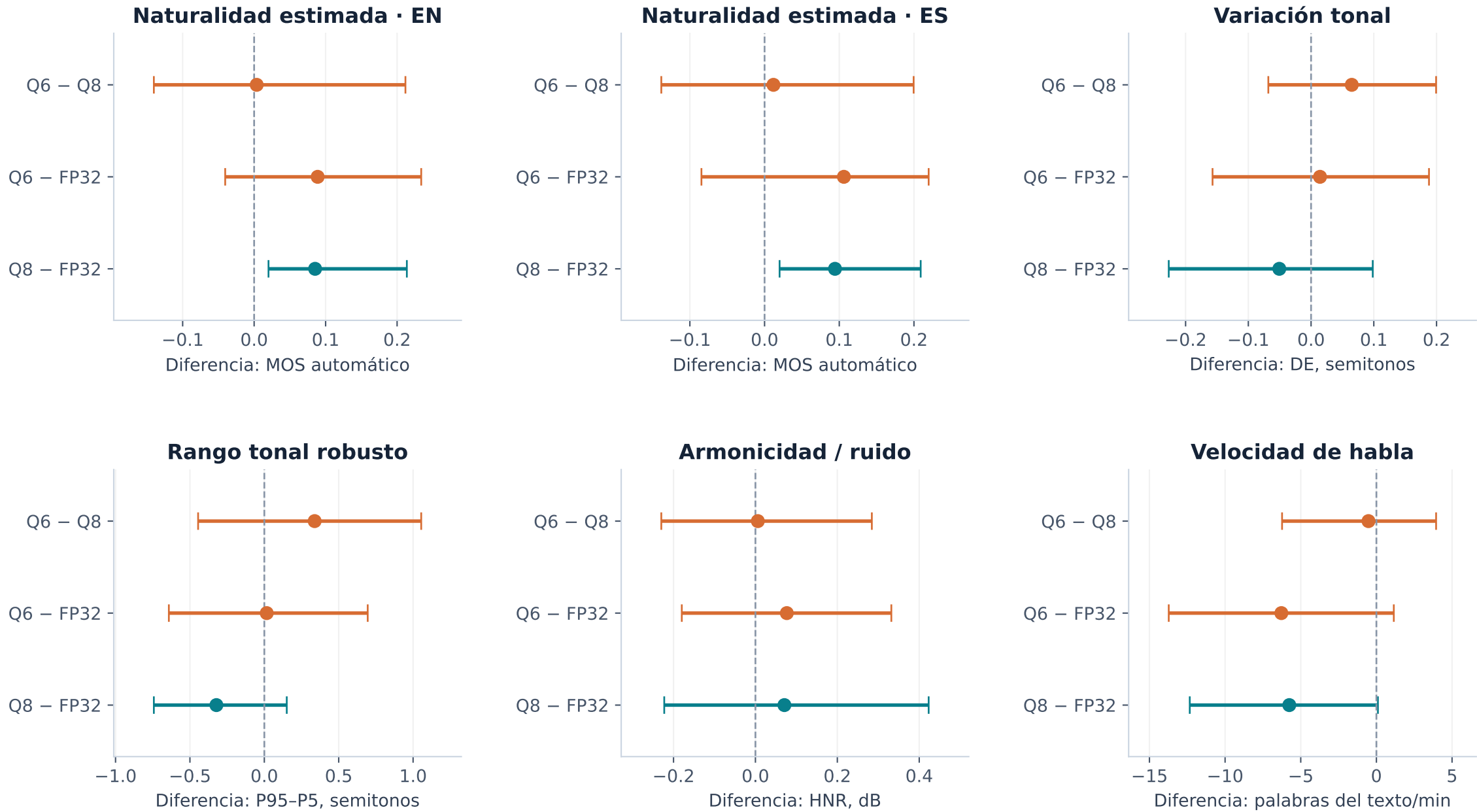




Jitter y shimmer en habla continua: descriptivos; no usar umbrales clínicos ni interpretarlos como naturalidad.
Acuerdo entre trackers: frames sonoros de ambos con diferencia ≤1 semitono. No mide exactitud frente a un F0 real conocido.

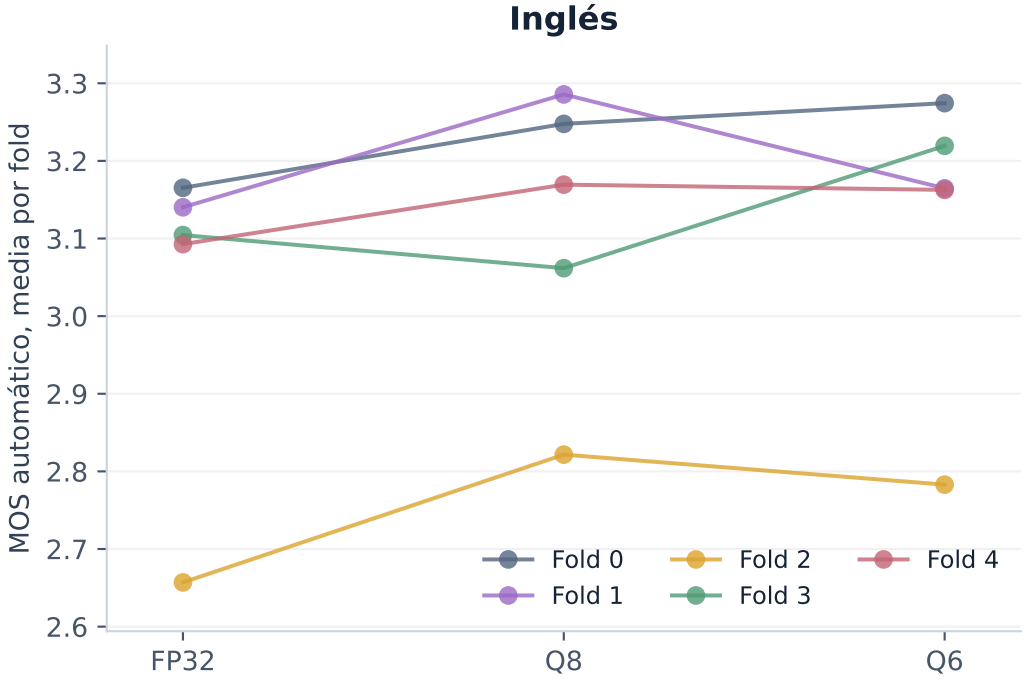
Diferencias pareadas y variabilidad entre frases

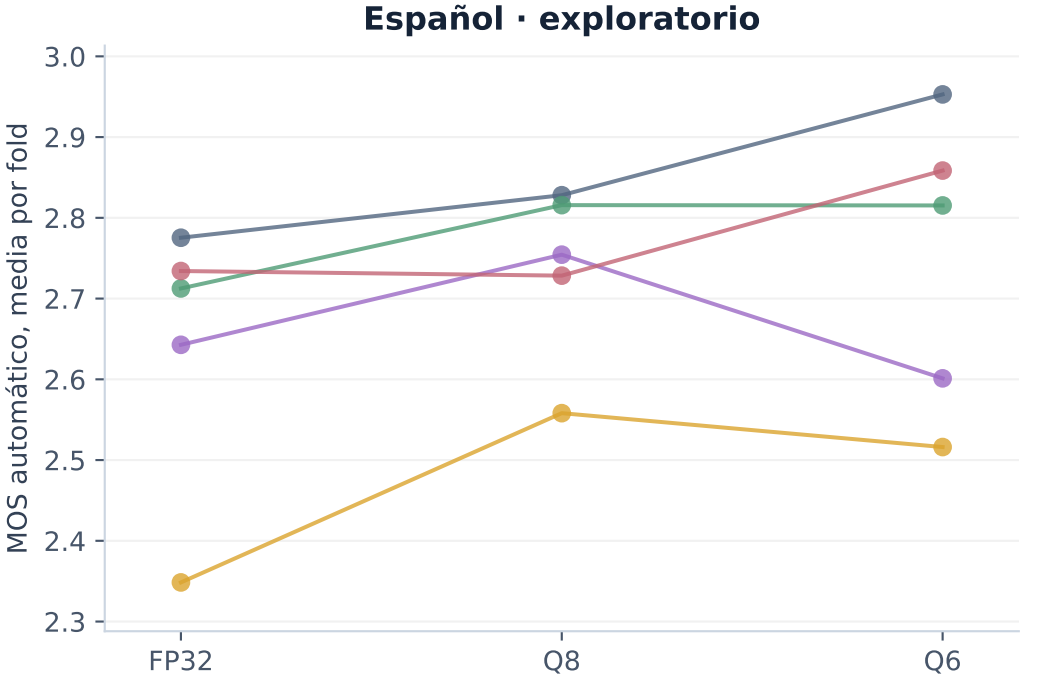
Cada diferencia compara la misma frase y semilla; se remuestrean frases, manteniendo ambas semillas



Barras: intervalo bootstrap descriptivo de 95%. Con 3 frases por idioma, la incertidumbre es amplia y parcial.
No incluye error de calibración de UTMOSv2, preferencias de oyentes ni variabilidad de voces o textos nuevos.

Naturalidad: variación entre los cinco evaluadores





Cada línea: un checkpoint UTMOSv2, cinco crops por clip. Su dispersión y los cambios de orden no son intervalos de MOS humano.
Los intervalos de las otras figuras solo remuestrean frases; no capturan esta variación del predictor.