MỤC LỤC

Chương 15. Bản quyền, đạo đức, phát hiện vi phạm 5
15.1 Vị trí của chương trong cuốn sách 6
15.2 Dấu vết codec — tệp này đã đi qua đâu 8
15.3 Thuỷ vân âm thanh 14
15.4 Vân tay âm thanh — nhận ra một bản ghi 21
15.5 Phát hiện giọng tổng hợp 26
15.6 Quyền, đạo đức, và chỗ kỹ thuật dừng lại 31
15.7 Bảo vệ sản phẩm của chính người đọc 33
15.8 Dự án thực hành — giai đoạn 5: hoàn thiện và trình bày 36
15.9 Tổng kết chương và tổng kết cuốn sách 39

DANH MỤC HÌNH VẼ

Hình 15.1 Bảy mục của chương 15.....6
Hình 15.2 Bốn câu hỏi và bốn công cụ tương ứng.....7
Hình 15.3 Phổ trung bình của cùng một bản nhạc qua ba đường.....9
Hình 15.4 Ba cách nhìn cùng một bảng: vách, lỗ phổ, và thứ hạng chất lượng.....10
Hình 15.5 Hai tệp cùng 322,6 kbit/giây, hai con số phổ hoàn toàn khác.....11
Hình 15.6 Hai dấu vết stereo và một dấu vết thời gian.....12
Hình 15.7 Ba sơ đồ và tam giác đánh đổi của chúng.....15
Hình 15.8 Thuỷ vân trải phổ trên dạng sóng, và thuỷ vân tiếng vọng trên cepstrum.....17
Hình 15.9 Tỉ lệ bit đọc sai của ba sơ đồ qua chín phép tấn công.....18
Hình 15.10 Hai phép đánh đổi, và cách chữa mất đồng bộ.....20
Hình 15.11 Năm bước từ phổ đồ tới một khoá tra cứu.....22
Hình 15.12 Hai chòm sao chồng lên nhau, và cách ghép cặp quanh một đỉnh neo.....23
Hình 15.13 Nhận dạng một đoạn trích ba giây dưới tám phép thử.....24
Hình 15.14 Độ dài đoạn trích cần thiết, và giá của một kho vân tay.....25
Hình 15.15 Cùng một câu nói, cùng một phổ mel — khác nhau ở PHA.....27
Hình 15.16 Phân bố độ nhọn, ngưỡng, và phép thử trên người nói chưa gặp.....28
Hình 15.17 Bốn đặc trưng, và máy dò sau khi tệp đã qua nén.....29
Hình 15.18 Sáu cách chữa, năm cách không qua được và một cách qua ngay.....30
Hình 15.19 Ba loại quyền trên cùng một bản ghi.....31
Hình 15.20 Năm câu hỏi phải trả lời trước khi phát hành một bản có giọng do AI sinh
.....32
Hình 15.21 Ba công cụ dưới bảy phép thử.....33
Hình 15.22 Thuỷ vân trước hay sau bộ giới hạn, và bộ kiểm bàn giao.....34
Hình 15.23 Năm kết quả cần có ở giai đoạn 5.....36
Hình 15.24 Các hạng mục tự đánh giá giai đoạn 5.....37
Hình 15.25 Chương 15 gói lại trong sáu con số.....40
Hình 15.26 Mười lăm chương trên một đường đi.....41

DANH MỤC BẢNG BIỂU

Bảng 15.1 Bốn công cụ và điều kiện dùng được của từng cái....7
Bảng 15.2 Ba loại vật liệu dùng suốt chương....8
Bảng 15.3 Ba loại dấu vết, và điều kiện đọc được từng loại....9
Bảng 15.4 Quét chín đường nén trên cùng một bản nhạc tám giây....10
Bảng 15.5 Bản gốc, bản 64k, bản chuyển mã, và bản 320k thật....11
Bảng 15.6 Stereo ghép đôi ăn vào kênh bên như thế nào....13
Bảng 15.7 Tiền vọng: năng lượng sai số ngay trước tiếng gõ....13
Bảng 15.8 Ba chiều của một sơ đồ thuỷ vân....14
Bảng 15.9 Tỉ lệ bit đọc sai — 0 là đọc đúng, 0,5 là đoán ngẫu nhiên....19
Bảng 15.10 Cắt đầu và cách chữa....20
Bảng 15.11 Quét mức nhúng — độ nghe thấy đổi lấy độ bền....21
Bảng 15.12 Quét độ dài khối — dung lượng đổi lấy độ bền....21
Bảng 15.13 Ba điều kiện mà một vân tay phải thoả....22
Bảng 15.14 Vì sao đỉnh phổ sống sót....22
Bảng 15.15 Truy vấn bằng một đoạn trích ba giây lấy từ giây thứ 2,5....24
Bảng 15.16 Đoạn trích lẫn nhiễu SNR 10 dB, theo độ dài....25
Bảng 15.17 Giá của một kho vân tay....25
Bảng 15.18 Bốn đặc trưng đã thử, trên cùng một cặp giọng....27
Bảng 15.19 Ngưỡng đặt trên một người nói, đem dùng cho ba người khác....28
Bảng 15.20 Độ nhọn sau bốn đường nén....29
Bảng 15.21 Sáu cách chữa và kết quả....30
Bảng 15.22 Ba lớp quyền và điều mỗi lớp bảo hộ....31
Bảng 15.23 Năm câu hỏi và điều phải làm với mỗi câu....32
Bảng 15.24 Ba công cụ trên cùng một bản nhạc đã nhúng thuỷ vân....33
Bảng 15.25 Hai vị trí, hai kết quả....34
Bảng 15.26 Bộ kiểm bàn giao của chương 14, trước và sau khi nhúng....35
Bảng 15.27 Thời gian máy cho mỗi giây âm thanh....35
Bảng 15.28 Năm giai đoạn của dự án thực hành....36
Bảng 15.29 Năm kết quả cần có và tiêu chuẩn của từng thứ....37
Bảng 15.30 Ba lỗi hay gặp nhất khi trình bày dự án....39
Bảng 15.31 Sáu con số và ý nghĩa của từng con số....40
Bảng 15.32 Năm thói quen đáng mang ra khỏi cuốn sách....41
Bảng 15.33 Năm hướng đi tiếp, và chỗ bắt đầu của mỗi hướng....42

CHƯƠNG 15
BẢN QUYỀN, ĐẠO ĐỨC, PHÁT HIỆN VI PHẠM

Mục tiêu của chương

Sau chương này, người đọc phải làm được những việc sau.

- Đọc được ba dấu vết codec trên một tệp lạ mà không cần bản gốc, và nói được tệp ấy đã đi qua những khâu nào.
- Chứng minh được bằng hai con số rằng một tệp bitrate cao thật ra được dựng lại từ một bản bitrate thấp.
- Giải thích được vì sao stereo ghép đôi làm ảnh stereo hẹp lại, và đo được mức độ hẹp ấy.
- Cài được ba sơ đồ thuỷ vân, và đặt được từng sơ đồ vào đúng chỗ của nó trong tam giác dung lượng – độ bền – độ nghe thấy.
- Đo được độ bền của một thuỷ vân bằng tỉ lệ bit sai qua một bộ phép tấn công, và chỉ ra được kẻ thù thật của nó.
- Dựng được vân tay âm thanh bằng chòm sao đỉnh phổ và khoá băm từng cặp, và đo được khả năng nhận dạng của nó dưới nhiễu.
- Giải thích được vì sao vân tay dựng trên *đỉnh* chứ không trên phổ, bằng con số chứ không bằng lập luận suông.
- Chọn được một đặc trưng phát hiện giọng tổng hợp, đặt ngưỡng cho nó bằng EER, và kiểm được ngưỡng ấy trên người nói chưa gặp.
- Chỉ ra được chỗ máy dò ấy mù, và nói rõ nó đang dò dấu vết gì chứ không phải dò "AI".
- Phân biệt được ba lớp quyền trên một bản ghi, và trả lời được năm câu hỏi trước khi phát hành một bản có giọng do AI sinh.
- Ghép được ba công cụ bảo vệ thành một chuỗi, đặt đúng vị trí trong đường ống của chương 14, và chứng minh nó không làm sai lệch bản giao.

Chương cuối cùng, và chương duy nhất có một mục không đo được

Bốn mục kỹ thuật của chương này — dấu vết codec, thuỷ vân, vân tay, phát hiện giọng tổng hợp — đều đo được đầy đủ, và mọi phép nén trong bài chạy bằng ffmpeg với chính bộ mã hoá mà người dùng thật dùng.

Mục 15.6 thì khác: nó không có một con số nào, và đó là chủ ý. Kỹ thuật nói chúng ta *làm được* gì; nó không nói chúng ta *nên* làm gì. Một cuốn sách kỹ thuật nghiêm túc phải chỉ ra được chỗ chính nó dừng lại.

Mục 15.5.5 khép lại phần kỹ thuật bằng một kết quả không dễ chấp nhận nhưng có thật: máy dò giọng tổng hợp tốt nhất của bài mù hoàn toàn trước một bộ dựng sóng biết dựng đúng pha. Đây là cuộc chạy đua, không phải một bài toán đã giải xong.

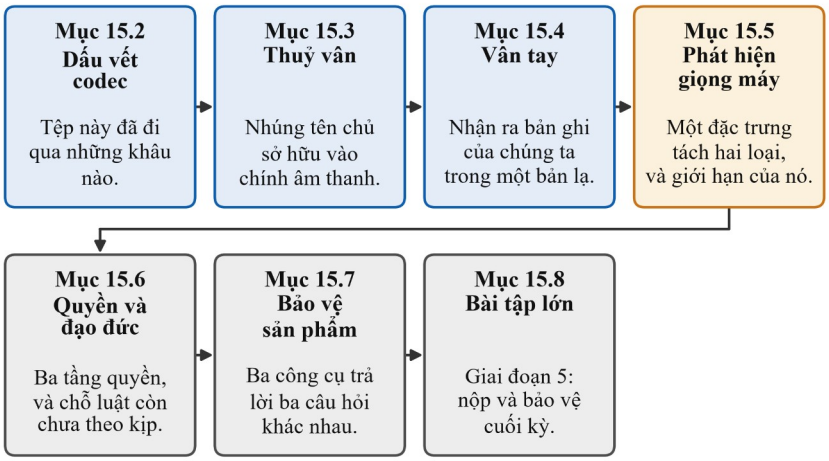
Mô-đun tính toán của cuốn sách khép lại với 480 phép tự kiểm, chạy bằng một lệnh, đạt 100 phần trăm.

15.1 Vị trí của chương trong cuốn sách

15.1.1 Mười bốn chương lo việc làm ra; chương này lo hai câu hỏi khác

Mười bốn chương vừa qua trả lời một câu hỏi duy nhất, mỗi chương một phần: làm ra âm thanh như thế nào. Chương cuối cùng đổi hẳn loại câu hỏi.

- **Câu thứ nhất — bản này ở đâu ra.** Một tệp đến tay người dùng đã đi qua những khâu nào, đã bị nén mấy lần, có phải bản gốc hay là bản dựng lại từ một bản kém hơn, và giọng trong đó là người hay là máy. Bốn mục kỹ thuật của chương này trả lời câu ấy bằng số đo.
- **Câu thứ hai — chúng ta được phép làm gì với nó.** Câu này *không* có phép đo. Nó cần luật, cần hợp đồng, và cần một quyết định của con người. Mục 15.6 nói thẳng về ranh giới ấy.

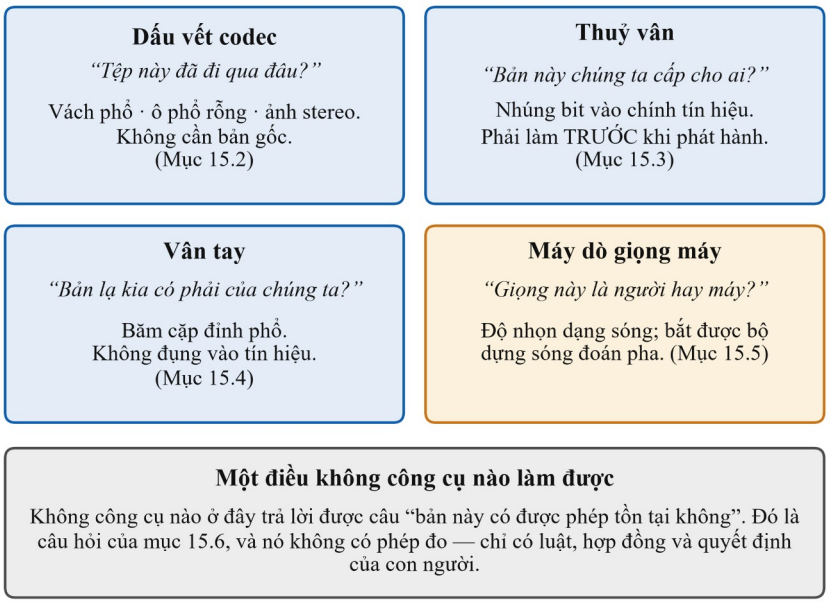


Mười bốn chương trước lo việc LÀM RA âm thanh. Chương cuối lo hai câu hỏi khác: bản này ở đâu ra, và chúng ta được phép làm gì với nó.

Hình 15.1 Bảy mục của chương 15

15.1.2 Bốn công cụ, bốn câu hỏi — không cái nào thay được cái nào

Điều hay bị lẫn nhất trong lĩnh vực này là gộp bốn công cụ khác nhau làm một. Chúng trả lời bốn câu hỏi khác nhau, và dùng nhầm công cụ thì không ra câu trả lời nào cả.



Hình 15.2 Bốn câu hỏi và bốn công cụ tương ứng

Bảng 15.1 Bốn công cụ và điều kiện dùng được của từng cái

Công cụ	Trả lời câu	Cần gì trước đó	Có đụng vào tín hiệu
Dấu vết codec	Tệp này đã đi qua đâu	Không cần gì cả	Không
Thuỷ vân	Bản này chúng ta cấp cho ai	Phải nhúng *trước* khi phát hành	Có — nhưng dưới ngưỡng nghe
Vân tay	Bản lạ kia có phải của chúng ta	Phải dựng kho từ bản gốc	Không
Máy dò giọng máy	Giọng này người hay máy	Phải hiệu chỉnh ngưỡng trên dữ liệu có nhãn	Không

Một điều không công cụ nào làm được

Không công cụ nào trong bảng trả lời được câu "bản này có được phép tồn tại không". Kỹ thuật nói chúng ta *làm được* gì; nó không nói chúng ta *nên* làm gì.

Đó không phải một câu nói cho hay. Nó là ranh giới thật, và mục 15.6 sẽ chỉ ra năm chỗ cụ thể mà ranh giới ấy đi qua.

15.1.3 Vật liệu và công cụ của chương

Chương này khác các chương trước ở một chỗ về công cụ: mọi phép nén trong bài đều chạy bằng ffmpeg với chính bộ mã hoá mà người dùng thật dùng — libmp3lame, aac, libopus, libvorbis — chứ không qua một mô hình giả lập.

Bảng 15.2 Ba loại vật liệu dùng suốt chương

Vật liệu	Dựng bằng gì	Dùng cho mục nào
Bản nhạc gốc 48 kHz	Giai điệu hài, lớp gõ, một chút khí ở vùng cao	15.2 · 15.3 · 15.4
Bản trộn stereo	Phần giữa chung, phần riêng khác nhau; tương quan hai kênh 0,9615	15.2.4
Giọng nói mô phỏng	Nguồn xung thanh môn có jitter và shimmer, lọc qua ba formant	15.5
Giọng do máy dựng lại	Chính giọng ấy đi qua phổ mel rồi Griffin-Lim	15.5

> **Vì sao dùng griffin-lim làm giọng máy"**
> Không thể chạy một mô hình sinh giọng thật trong một chương, và cũng không cần. Điều mục 15.5 đo là dấu vết mà *khâu dựng lại sóng* để lại — và Griffin-Lim để lại đúng loại dấu vết ấy, vì nó cũng phải đoán pha từ độ lớn phổ.
>
> Đổi lại, phải nói rõ giới hạn: một bộ dựng sóng học được hiện đại dựng pha tốt hơn Griffin-Lim rất nhiều. Mục 15.5.5 đo thẳng điều đó, và kết quả là máy dò của bài mù hoàn toàn trước nó.

15.2 Dấu vết codec — tệp này đã đi qua đâu

15.2.1 Vì sao một bộ mã hoá có mất mát luôn để lại dấu vết

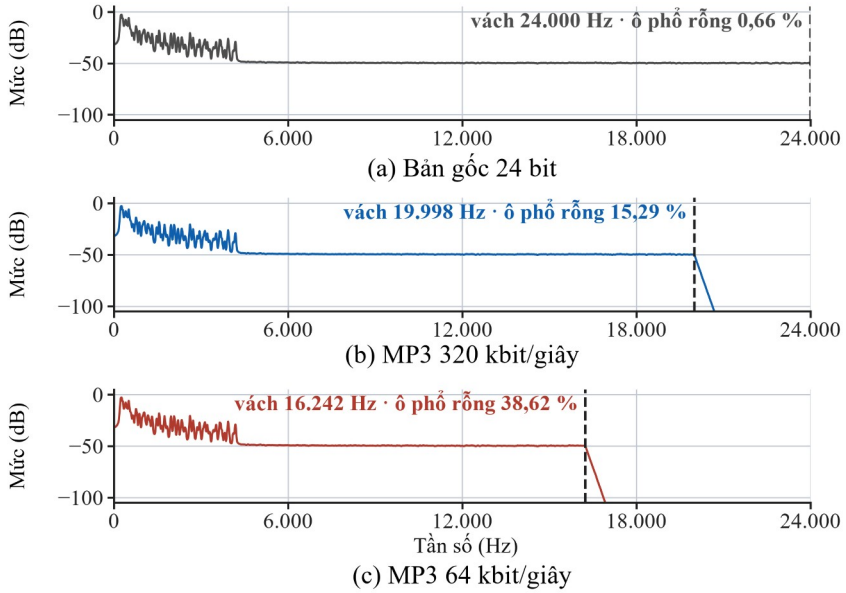
Mọi bộ mã hoá có mất mát đều làm cùng một việc: chia phổ thành các dải, tính xem tai nghe được bao nhiêu ở mỗi dải, rồi cấp bit theo đó. Dải nào tai không nghe thì cấp không bit — nghĩa là loại bỏ hẳn.

Chính phép "loại bỏ hẳn" ấy để lại dấu vết. Âm thanh tự nhiên không bao giờ có những ô phổ rỗng tuyệt đối, và không bao giờ tắt hẳn từ một tần số trở lên. Tệp đã qua nén thì có cả hai.

Bảng 15.3 Ba loại dấu vết, và điều kiện đọc được từng loại

Dấu vết	Đọc được khi nào	Bền tới đâu
Vách phổ	Luôn luôn — chỉ cần chính tệp ấy	Sống qua mọi phép xử lý sau đó, kể cả nén lại
Tỉ lệ ô phổ rỗng	Luôn luôn	Bền nhất trong ba loại; không phụ thuộc tần số lấy mẫu
Ảnh stereo hẹp bất thường	Khi tệp có hai kênh	Bền, nhưng cần biết bản gốc rộng bao nhiêu mới chắc
Tiền vọng quanh tiếng gõ	Chỉ khi có bản gốc để so	Yếu — mọi phép xử lý sau đó đều làm nhoè

15.2.2 Vách phổ và tỉ lệ ô phổ rỗng



Hình 15.3 Phổ trung bình của cùng một bản nhạc qua ba đường

Vách phổ dò bằng đúng hàm đã dựng ở mục 14.2.5. Dấu vết thứ hai thì mới, và nó đơn giản hơn: đếm xem bao nhiêu phần trăm số ô của phổ đồ nằm dưới đỉnh một khoảng lớn.

$$L_{\text{rỗng}} = \frac{100}{TK}\left|\{(t,k):\ 20\log_{10}|X(t,k)| < X_{\max} - \Delta\}\right| \tag{15.1}$$

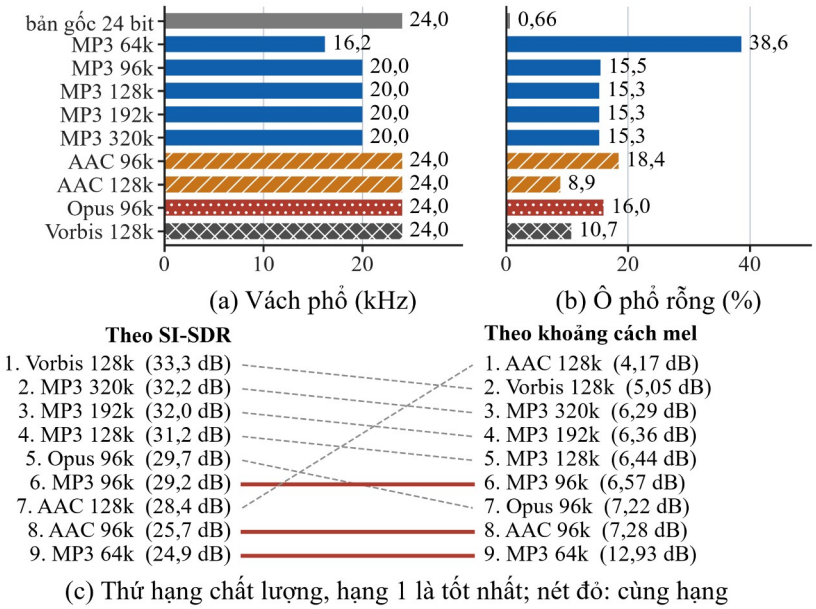
Ở đây T là số khung, K là số vạch tần số của mỗi khung, và delta là khoảng cách tính từ ô lớn nhất của cả phổ đồ. Bài dùng delta bằng 90 decibel — đủ sâu để không đếm nhầm những chỗ vốn yên.

Liệt kê 15.1 Đếm tỉ lệ ô phổ bị đưa về 0

```
def lo_pho_pt(x, nguong_db=-90.0, n_fft=2048, hop=512):
    X, _, _, _ = _stft(sang_mono(x), n_fft, hop)
    db = 20.0 * np.log10(np.maximum(np.abs(X), 1e-15))
    return float(100.0 * np.mean(db - db.max() < nguong_db))
```

Bảng 15.4 Quét chín đường nén trên cùng một bản nhạc tám giây

Bản	Vách phổ	Ô phổ rỗng	SI-SDR	Khoảng cách mel	Tiền vọng
bản gốc 24 bit	24.000 Hz	0,66 %	—	—	—
MP3 64k	16.242 Hz	38,62 %	24,9 dB	12,93 dB	7,3 dB
MP3 96k	19.998 Hz	15,50 %	29,2 dB	6,57 dB	6,7 dB
MP3 128k	19.992 Hz	15,30 %	31,2 dB	6,44 dB	5,7 dB
MP3 192k	19.998 Hz	15,26 %	32,0 dB	6,36 dB	4,3 dB
MP3 320k	19.998 Hz	15,29 %	32,2 dB	6,29 dB	0,6 dB
AAC 96k	24.000 Hz	18,44 %	25,7 dB	7,28 dB	12,0 dB
AAC 128k	24.000 Hz	8,86 %	28,4 dB	4,17 dB	14,8 dB
Opus 96k	24.000 Hz	16,00 %	29,7 dB	7,22 dB	2,2 dB
Vorbis 128k	24.000 Hz	10,68 %	33,3 dB	5,05 dB	2,4 dB



Hình 15.4 Ba cách nhìn cùng một bảng: vách, lỗ phổ, và thứ hạng chất lượng

Ba điều đọc được từ bảng này

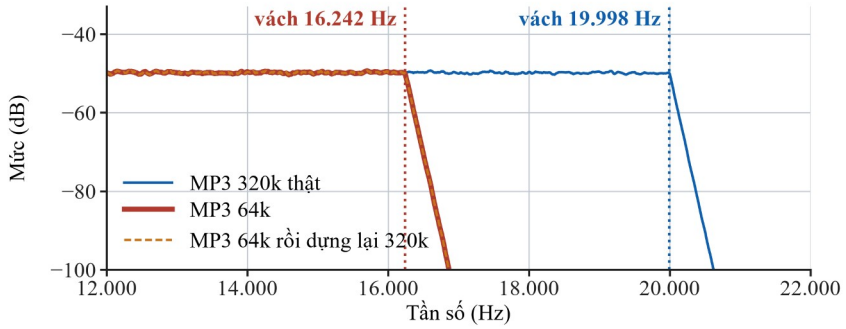
MP3 *cắt*, các codec khác thì không. Lame mặc định cắt ở khoảng 20 kHz từ 96 kbit/giây trở lên, và ở 16,2 kHz tại 64 kbit/giây. AAC, Opus và Vorbis giữ tới Nyquist. Vì thế "có vách ở 16 kHz" gần như luôn nghĩa là MP3 bitrate thấp.

Tỉ lệ ô phổ rỗng tách bản gốc khỏi mọi bản nén rất dứt khoát: 0,66 % so với 8,86 % của bản nén sạch nhất trong bảng. Không có chỗ nào chồng lấn.

Hai thước đo chất lượng cho *hai* thứ hạng khác nhau — chỉ 2 trong 9 hạng trùng nhau. SI-SDR phạt sai lệch dạng sóng; khoảng cách mel phạt sai lệch mà tai nghe được. Codec tối ưu cho cái thứ hai chứ không cho cái thứ nhất.

15.2.3 Chuyển mã: chỗ bitrate ghi trong tệp không đúng thực tế

Đây là trường hợp gặp nhiều nhất trong thực tế: ai đó lấy một bản 64 kbit/giây, giải nén rồi nén lại ở 320 kbit/giây, và đưa cho chúng ta như một bản chất lượng cao.



(a) Vách phổ không đổi khi dựng lại ở bitrate cao

Bản	Vách phổ	Ô phổ rỗng	Bitrate thật
bản gốc	24.000 Hz	0,66 %	1.152 kbit/giây
MP3 64k	16.242 Hz	38,62 %	64,6 kbit/giây
MP3 64k rồi dựng lại 320k	16.242 Hz	38,61 %	322,6 kbit/giây
MP3 320k thật	19.998 Hz	15,29 %	322,6 kbit/giây

(b) Hai tệp cùng 322,6 kbit/giây, hai con số phổ khác nhau

Hình 15.5 Hai tệp cùng 322,6 kbit/giây, hai con số phổ hoàn toàn khác

Bảng 15.5 Bản gốc, bản 64k, bản chuyển mã, và bản 320k thật

Bản	Vách phổ	Ô phổ rỗng	Bitrate thật
bản gốc	24.000 Hz	0,66 %	1.152 kbit/giây
MP3 64k	16.242 Hz	38,62 %	64,6 kbit/giây
MP3 64k rồi dựng lại 320k	16.242 Hz	38,61 %	322,6 kbit/giây
MP3 320k thật	19.998 Hz	15,29 %	322,6 kbit/giây

Hai dòng cuối là toàn bộ mục này

Hai tệp có cùng một bitrate, 322,6 kbit/giây. Nhưng một tệp có vách ở 16.242 Hz và 38,61 % ô rỗng, còn tệp kia có vách ở 19.998 Hz và 15,29 %. Chênh lệch không nhỏ đi một chút nào sau khi nén lại ở bitrate cao.

Lý do đơn giản: nén lại *không* thêm được thông tin. Bộ mã hoá thứ hai chỉ nhận được thứ mà bộ thứ nhất còn chừa lại, và nó mã đúng thứ ấy thật kỹ.

Hệ quả thực hành: đừng bao giờ đánh giá chất lượng một tệp bằng bitrate ghi trong đó. Hai lệnh đo — vách phổ và tỉ lệ ô rỗng — nói ra sự thật trong vài giây.

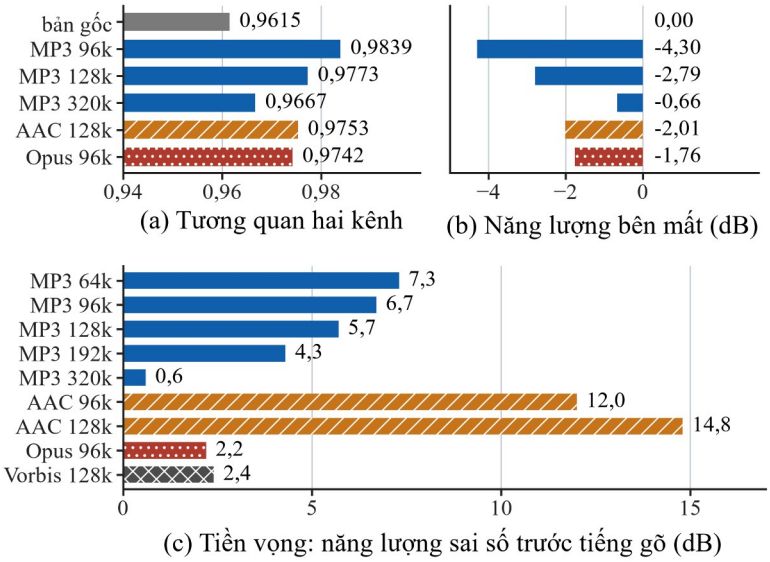
15.2.4 Stereo ghép đôi và tiền vọng

Hai dấu vết còn lại yếu hơn nhưng đáng biết, vì chúng chỉ ra hai kỹ thuật mà mọi bộ mã hoá đều dùng.

Kỹ thuật thứ nhất là stereo ghép đôi. Thay vì mã hai kênh trái và phải riêng, bộ mã hoá đổi sang kênh *giữa* và kênh *bên*.

$$M = \frac{L+R}{2}, \qquad S = \frac{L-R}{2}, \qquad \eta = 10\log_{10}\frac{\langle S^2\rangle}{\langle M^2\rangle} \tag{15.2}$$

Với một bản trộn thường, kênh giữa mang gần hết năng lượng còn kênh bên rất nhỏ — ở vật liệu của bài là −17,08 decibel. Bộ mã hoá vì thế cấp gần hết bit cho kênh giữa, và kênh bên là thứ bị loại bỏ trước tiên khi thiếu bit.



Hình 15.6 Hai dấu vết stereo và một dấu vết thời gian

Bảng 15.6 Stereo ghép đôi ăn vào kênh bên như thế nào

Bản	Tương quan hai kênh	Bên trên giữa	Năng lượng bên mất	SI-SDR giữa	SI-SDR bên
bản gốc	0,9615	−17,08 dB	—	—	—
MP3 96k	0,9839	−20,90 dB	−4,30 dB	18,2 dB	−5,3 dB
MP3 128k	0,9773	−19,39 dB	−2,79 dB	19,9 dB	−2,9 dB
MP3 320k	0,9667	−17,72 dB	−0,66 dB	23,8 dB	6,3 dB
AAC 128k	0,9753	−19,03 dB	−2,01 dB	17,9 dB	−3,1 dB
Opus 96k	0,9742	−18,84 dB	−1,76 dB	18,0 dB	−1,2 dB

Hai cột cuối là chỗ đáng dừng lại. Ở 96 kbit/giây, kênh giữa đạt SI-SDR 18,2 decibel — tốt — trong khi kênh bên đạt *âm* 5,3 decibel, nghĩa là phần sai còn lớn hơn phần đúng. Bộ mã hoá đã gần như loại bỏ hẳn kênh bên và dựng lại một thứ khác.

> **Hệ quả nghe được, và hệ quả pháp lý**
>
> Hệ quả nghe được: ảnh stereo hẹp lại. Tương quan hai kênh đi từ 0,9615 lên 0,9839 — bản nhạc nghe gần mono hơn. Đây là thứ người nghe mô tả là "bẹt" mà không chỉ ra được vì sao.
>
> Hệ quả pháp lý: một tệp tự nhận là bản thu stereo gốc mà có tương quan hai kênh cao bất thường là dấu hiệu đáng nghi. Nó không phải bằng chứng — một bản trộn vốn đã hẹp cũng cho con số ấy — nhưng nó là lý do để đo tiếp.

Kỹ thuật thứ hai là cửa sổ dài. Bộ mã hoá làm việc trên từng khối vài trăm tới hơn nghìn mẫu, và sai số lượng tử trải đều trong cả khối. Khi trong khối có một tiếng gõ, phần sai số ấy rò ra cả phần *trước* tiếng gõ — chỗ đáng lẽ vẫn yên. Hiện tượng ấy gọi là tiền vọng.

Bảng 15.7 Tiền vọng: năng lượng sai số ngay trước tiếng gõ

Bản	Tiền vọng	Nhận xét
MP3 64k	7,3 dB	Nghe thấy rõ nhất, như tiếng rít nhẹ trước nhịp
MP3 128k	5,7 dB	Giảm theo bitrate
MP3 320k	0,6 dB	Gần như hết
AAC 128k	14,8 dB	Cao hơn MP3 vì cửa sổ dài hơn
Opus 96k	2,2 dB	Thấp — Opus đổi cửa sổ rất linh hoạt
Vorbis 128k	2,4 dB	Thấp, cùng lý do

Một phép đo đã thử và không dùng được

Mọi bộ mã hoá đều làm việc trên một *lưới* khung cố định — MP3 dùng khối 576 mẫu, AAC dùng 1.024. Về lý thuyết, lưới ấy phải để lại dấu vết tuần hoàn trong đường bao năng lượng cao tần, và dò được bằng cách tìm đỉnh phổ tại tần số bằng tần số lấy mẫu chia cho độ dài khung.

Bài đã thử phép đo ấy và nó *không* dùng được trên vật liệu này: chính bản gốc đã có một thành phần mạnh đúng tại tần số ứng với 576 mẫu, nên bản gốc và bản MP3 cho con số gần như nhau. Phép đo đang đo nội dung nhạc chứ không đo lưới khung.

Ghi lại ở đây vì đó là một phần của cách làm việc: phép đo nào đã thử và không dùng được thì nói rõ, chứ không âm thầm bỏ đi. Hai phép đo dùng được — vách phổ và tỉ lệ ô rỗng — đã đủ mạnh.

15.3 Thuỷ vân âm thanh

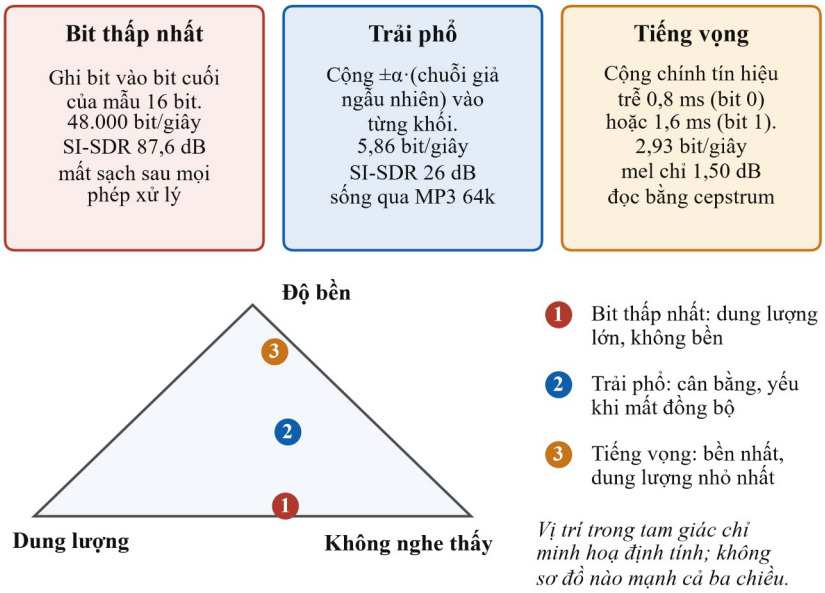
15.3.1 Ba chiều không thể cùng tốt

Thuỷ vân là việc giấu một chuỗi bit vào chính tín hiệu, sao cho tai không nghe thấy nhưng máy đọc lại được. Nó trả lời một câu hỏi rất cụ thể: bản này chúng ta cấp cho ai.

Mọi sơ đồ thuỷ vân đều phải cân ba chiều, và không sơ đồ nào mạnh ở cả ba.

Bảng 15.8 Ba chiều của một sơ đồ thuỷ vân

Chiều	Nghĩa là gì	Đo bằng gì
Dung lượng	Giấu được bao nhiêu bit mỗi giây	Số bit trên tổng số giây
Độ bền	Còn đọc được sau những phép xử lý nào	Tỉ lệ bit đọc sai sau mỗi phép tấn công
Độ nghe thấy	Nghe ra được hay không	SI-SDR và khoảng cách mel — và hai con số ấy hay bất đồng



Hình 15.7 Ba sơ đồ và tam giác đánh đổi của chúng

15.3.2 Sơ đồ thứ nhất: bit thấp nhất — và vì sao nó vô dụng

Cách đơn giản nhất là ghi bit vào bit thấp nhất của mỗi mẫu đã lượng tử hoá. Dung lượng bằng đúng tần số lấy mẫu: 48.000 bit mỗi giây, nhiều hơn hai sơ đồ kia hàng vạn lần.

Liệt kê 15.2 Nhúng và đọc thuỷ vân bit thấp nhất

```
def thuy_van_lsb_nhung(x, bit, so_bit_mau=16):
    q = np.round(np.clip(x, -1, 1) * (2**(so_bit_mau-1) - 1))
    q = q.astype(np.int64)
    q[:len(bit)] = (q[:len(bit)] & ~1) | np.asarray(bit)
    return q.astype(float) / (2**(so_bit_mau-1) - 1)
```

Nó cũng hoàn toàn không nghe thấy: SI-SDR đo được 87,6 decibel, và khoảng cách mel 0,01 decibel. Về hai chiều đầu thì không sơ đồ nào bằng.

Và nó chết vì mọi thứ

Giảm mức 6 decibel: tỉ lệ bit đọc sai 0,485. Đổi tần số lấy mẫu: 0,482. Nén MP3 dù ở 128 hay 320 kbit/giây: khoảng 0,49.

Con số 0,5 nghĩa là đoán ngẫu nhiên. Nghĩa là sau bất cứ phép xử lý nào, thuỷ vân bit thấp nhất không còn lại gì cả — nó không "yếu đi" mà biến mất hoàn toàn.

Lý do hiển nhiên khi nói ra: bit thấp nhất là thứ mang *ít* thông tin nghe được nhất, nên nó cũng là thứ mọi phép xử lý phá trước tiên. Chính điều làm nó không nghe thấy cũng là điều làm nó không bền.

Sơ đồ này vẫn có chỗ dùng: kiểm toàn vẹn trong một hệ thống khép kín, nơi tệp không đi qua phép xử lý nào. Ngoài chỗ ấy thì không.

15.3.3 Sơ đồ thứ hai: trải phổ

Ý tưởng đến từ truyền thông: thay vì giấu bit vào một chỗ nhỏ, hãy trải nó ra rất nhiều mẫu với biên độ rất thấp, rồi đọc lại bằng tương quan.

$$y[n] = x[n] + (2b - 1)\,\alpha\,\sigma_{\text{khối}}\,p[n], \qquad \alpha = 10^{G/20} \tag{15.3}$$

Mỗi khối N mẫu mang một bit. Ta cộng vào khối ấy một chuỗi giả ngẫu nhiên p, với dấu tuỳ theo bit, và biên độ tỉ lệ với mức hiệu dụng của *chính* khối ấy — nên chỗ nhạc to mang thuỷ vân mạnh và chỗ yên mang thuỷ vân yếu, đúng theo đường che lấp của mục 1.7.

$$\hat{b} = \frac{1}{2}\left[1 + \operatorname{sgn}\sum_{n=0}^{N-1} y[n]\,p[n]\right] \tag{15.4}$$

Đọc lại chỉ là một phép nhân vô hướng. Nếu bit là 1 thì tổng dương, nếu là 0 thì âm; chuỗi p đã biết trước nên không cần bản gốc.

Liệt kê 15.3 Nhúng thuỷ vân trải phổ

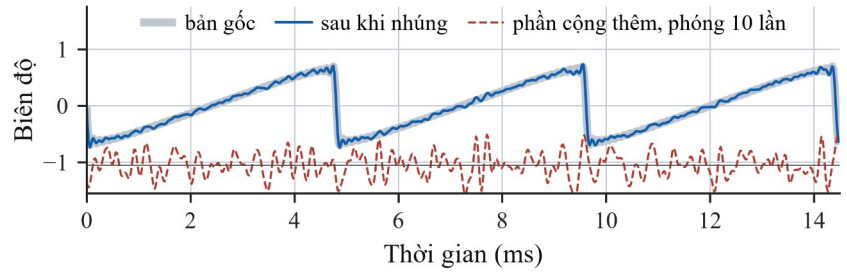
```
def thuy_van_trai_pho_nhung(x, bit, sr, chip=8192, muc_db=-26.0,
                            hat=0):
    y = x.copy()
    p = chuoi_pn(chip, hat, sr)          # da loc ve 500 – 5000 Hz
    g = db_to_lin(muc_db)
    for k, b in enumerate(bit):
        i = k * chip
        seg = y[i:i+chip]
        a = g * np.sqrt(np.mean(seg**2))
        y[i:i+chip] = seg + (1.0 if b else -1.0) * a * p
    return y
```

Một chi tiết quyết định độ bền

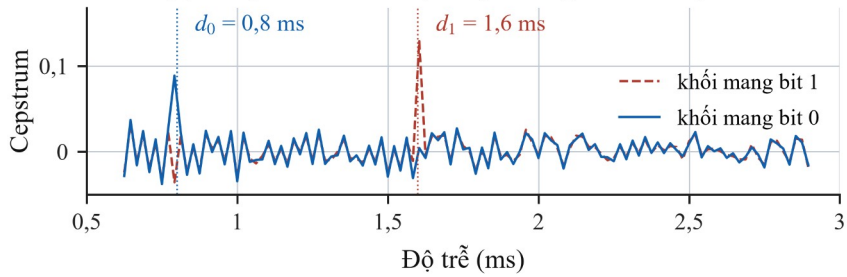
Chuỗi giả ngẫu nhiên *không* để trải đều tới Nyquist mà lọc về dải 500 đến 5.000 Hz trước khi dùng.

Lý do là mục 15.2: codec cắt phần trên vách phổ và loại bỏ những ô nhỏ ở vùng cao. Nếu thuỷ vân nằm ở đó thì nó mất ngay ở bước nén đầu tiên. Đặt nó vào đúng dải mà codec buộc phải giữ — dải tai nghe rõ nhất — thì nó sống.

Đây là ví dụ đẹp cho một nguyên tắc chung: muốn giấu một thứ bền, hãy giấu nó vào chỗ mà kẻ phá *không* dám đụng vào.



(a) Trải phổ: SI-SDR 26,0 dB, khoảng cách mel 8,35 dB



(b) Tiếng vọng: SI-SDR 11,8 dB, khoảng cách mel 1,50 dB

Hình 15.8 Thuỷ vân trải phổ trên dạng sóng, và thuỷ vân tiếng vọng trên cepstrum

$$\mathrm{BER} = \frac{1}{N_{\mathrm{bit}}} \sum_{i=1}^{N_{\mathrm{bit}}} [\hat{b}_i \neq b_i] \tag{15.5}$$

Thước đo độ bền là tỉ lệ bit đọc sai. Giá trị 0 nghĩa là đọc đúng hoàn toàn; giá trị 0,5 nghĩa là đoán ngẫu nhiên, tức là thuỷ vân đã mất sạch.

15.3.4 Sơ đồ thứ ba: ẩn bằng tiếng vọng

Sơ đồ thứ ba không cộng thêm thứ gì lạ vào tín hiệu. Nó cộng *chính* tín hiệu, bị trễ đi một khoảng rất ngắn — và bit nằm ở chỗ chọn khoảng trễ nào.

$$y[n] = x[n] + a\,x[n - d_b], \qquad c[m] = \mathsf{F}^{-1}\{\log|Y(f)|\} \tag{15.6}$$

Khoảng trễ d nhận một trong hai giá trị: 0,8 mili giây cho bit 0 và 1,6 mili giây cho bit 1. Đọc lại bằng cepstrum — biến đổi Fourier ngược của loga phổ — vì một tiếng vọng ở độ trễ d làm phổ loga gợn đều với chu kỳ 1/d, và cepstrum có một đỉnh đúng tại d.

Liệt kê 15.4 Đọc thuỷ vân tiếng vọng bằng cepstrum

```
def thuy_van_vong_doc(y, so_bit, sr, d0=0.0008, d1=0.0016,
                      khoi=16384):
    n0, n1 = int(d0*sr), int(d1*sr)
    out = []
    for k in range(so_bit):
        seg = y[k*khoi:(k+1)*khoi] * np.hanning(khoi)
        X = np.fft.rfft(seg)
        c = np.fft.irfft(np.log(np.maximum(np.abs(X), 1e-12)))
        out.append(1 if c[n1] > c[n0] else 0)
    return np.array(out)
```

Chỗ hai thước đo bất đồng nặng nhất trong cả cuốn sách

SI-SDR của bản có thuỷ vân tiếng vọng chỉ 11,8 decibel. Đọc con số ấy theo thói quen thì kết luận là sai nghiêm trọng.

Khoảng cách mel của chính bản ấy là 1,50 decibel — nhỏ hơn cả bản có thuỷ vân trải phổ, vốn có SI-SDR 26 decibel.

Lý do: thứ cộng thêm không phải nhiễu mà là *chính* tín hiệu bị trễ. Tai nghe nó là tiếng phòng, không phải tiếng tạp. SI-SDR đo sai lệch dạng sóng nên phạt rất nặng; tai thì không.

Đây là lần thứ năm trong cuốn sách một thước đo khách quan nói ra điều không đúng với tai — ba lần ở mục 13.8, một lần ở mục 14.6.2, và lần này.

15.3.5 Ba sơ đồ dưới chín phép tấn công

Phép tấn công	Bit thấp nhất	Trải phổ	Tiếng vọng
không tác động	0,000	0,000	0,000
giảm mức 6 dB	0,485	0,000	0,000
cộng nhiễu −30 dB	0,499	0,000	0,000
đổi 48 → 44,1 → 48 kHz	0,482	0,000	0,000
thông thấp 8 kHz	0,502	0,087	0,000
MP3 128 kbit/giây	0,496	0,000	0,000
MP3 64 kbit/giây	0,501	0,000	0,043
AAC 128 kbit/giây	0,493	0,000	0,000
cắt 1.000 mẫu ở đầu	0,499	0,696	0,000

đọc gần như không sai (không quá 0,01)
sai một phần (dưới 0,1)
từ 0,48 trở lên: thuỷ vân không còn gì (0,5 là đoán ngẫu nhiên)

Hình 15.9 Tỉ lệ bit đọc sai của ba sơ đồ qua chín phép tấn công

Bảng 15.9 Tỉ lệ bit đọc sai — 0 là đọc đúng, 0,5 là đoán ngẫu nhiên

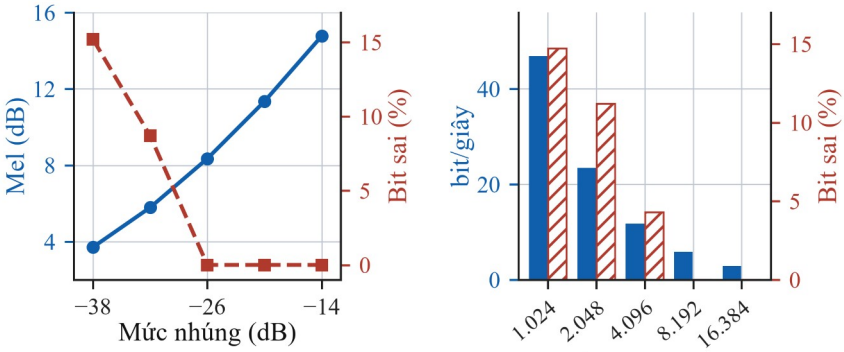
Phép tấn công	Bit thấp nhất	Trải phổ	Tiếng vọng
không đụng gì	0,000	0,000	0,000
giảm mức 6 dB	0,485	0,000	0,000
cộng nhiễu −30 dB	0,499	0,000	0,000
đổi 48 → 44,1 → 48 kHz	0,482	0,000	0,000
thông thấp 8 kHz	0,502	0,087	0,000
MP3 128 kbit/giây	0,496	0,000	0,000
MP3 64 kbit/giây	0,501	0,000	0,043
AAC 128 kbit/giây	0,493	0,000	0,000
cắt 1.000 mẫu ở đầu	0,499	0,696	0,000
Dung lượng (bit/giây)	48.000	5,86	2,93

Hàng cuối là cái giá. Sơ đồ bền nhất có dung lượng nhỏ hơn sơ đồ yếu nhất mười sáu nghìn lần. Với việc cần làm ở đây — nhúng một mã định danh vài chục bit vào một bản nhạc vài phút — thì 2,93 bit mỗi giây là quá đủ.

15.3.6 Kẻ thù thật của trải phổ: mất đồng bộ

Hàng áp chót của bảng trên là hàng đáng chú ý nhất. Thuỷ vân trải phổ sống sót qua MP3 64 kbit/giây với tỉ lệ bit sai bằng 0 — nhưng chết vì một phép cắt 1.000 mẫu, tức 21 mili giây.

Lý do không phải tín hiệu bị lỗi mà là mọi khối bị lệch chỗ. Phép đọc lấy khối thứ k từ mẫu k·N, và nếu đầu tệp mất 1.000 mẫu thì mỗi khối đọc ra là một phần của hai khối nhúng.



(a) Quét mức nhúng
(bit sai: sau MP3 64 kbit/giây)

(b) Quét độ dài khối
(bit sai: sau MP3 64 kbit/giây)

Cắt 1.000 mẫu (21 ms) ở đầu

đọc thẳng: bit sai 69,6 %
(mọi khối bị lệch chỗ)

Tìm lại đồng bộ rồi đọc

độ lệch tìm được 7.192 mẫu
(đúng 7.192); bit sai 0,0 %

(c) Mất đồng bộ và cách chữa

Hình 15.10 Hai phép đánh đổi, và cách chữa mất đồng bộ

Liệt kê 15.5 Tìm lại đồng bộ bằng chính chuỗi giả ngẫu nhiên

```
def tim_dong_bo_trai_pho(y, sr, chip=8192, hat=0):
    p = chuoi_pn(chip, hat, sr)
    tot, best = -1.0, 0
    for d in range(chip):
        v = sum(abs(np.dot(y[d+k*chip : d+(k+1)*chip], p))
                for k in range(4))
        if v > tot:
            tot, best = v, d
    return best
```

Bảng 15.10 Cắt đầu và cách chữa

	Giá trị
Số mẫu bị cắt	1.000
Tỉ lệ bit sai nếu đọc thẳng	0,696
Độ lệch tìm được	7.192 mẫu
Độ lệch đúng	7.192 mẫu
Tỉ lệ bit sai sau khi đồng bộ lại	0,000

Vì sao hệ thống thật luôn có mẫu đồng bộ

Phép tìm đồng bộ ở trên quét toàn bộ 8.192 độ lệch có thể, mỗi độ lệch tính bốn tích vô hướng. Nó chạy được, nhưng tốn.

Hệ thống thật giải bài toán ấy bằng cách nhúng thêm một *mẫu đồng bộ* — một chuỗi biết trước, đặt đều đặn — rồi tìm nó bằng tương quan. Cách ấy nhanh hơn và chịu được cả trường hợp tệp bị cắt ở giữa chứ không chỉ ở đầu.

Bài tập 3 yêu cầu người đọc thêm mẫu đồng bộ vào sơ đồ trải phổ và đo lại hàng cuối của bảng ở mục 15.3.5.

15.3.7 Hai phép đánh đổi, đo bằng số

Bảng 15.11 Quét mức nhúng — độ nghe thấy đổi lấy độ bền

Mức nhúng	SI-SDR	Khoảng cách mel	Bit sai (bản sạch)	Bit sai (sau MP3 64k)
−14 dB	14,0 dB	14,77 dB	0,000	0,000
−20 dB	20,0 dB	11,35 dB	0,000	0,000
−26 dB	26,0 dB	8,35 dB	0,000	0,000
−32 dB	32,0 dB	5,79 dB	0,087	0,087
−38 dB	38,0 dB	3,71 dB	0,130	0,152

Bảng 15.12 Quét độ dài khối — dung lượng đổi lấy độ bền

Số mẫu mỗi bit	Dung lượng	Bit sai (sạch)	Bit sai (MP3 128k)	Bit sai (MP3 64k)
1.024	46,88 bit/giây	0,144	0,144	0,147
2.048	23,44 bit/giây	0,107	0,107	0,112
4.096	11,72 bit/giây	0,022	0,032	0,043
8.192	5,86 bit/giây	0,000	0,000	0,000
16.384	2,93 bit/giây	0,000	0,000	0,000

Hai bảng cho cùng một hình dạng: có một ngưỡng, và vượt qua ngưỡng ấy thì sai rất nhanh. Với vật liệu của bài, mức −26 decibel và khối 8.192 mẫu là chỗ an toàn — và đó là cấu hình dùng ở mục 15.7.

15.4 Vân tay âm thanh — nhận ra một bản ghi

15.4.1 Vân tay khác thuỷ vân ở chỗ nào

Thuỷ vân phải nhúng *trước* khi phát hành. Nếu bản nhạc đã ra ngoài mà không có thuỷ vân thì không làm lại được nữa.

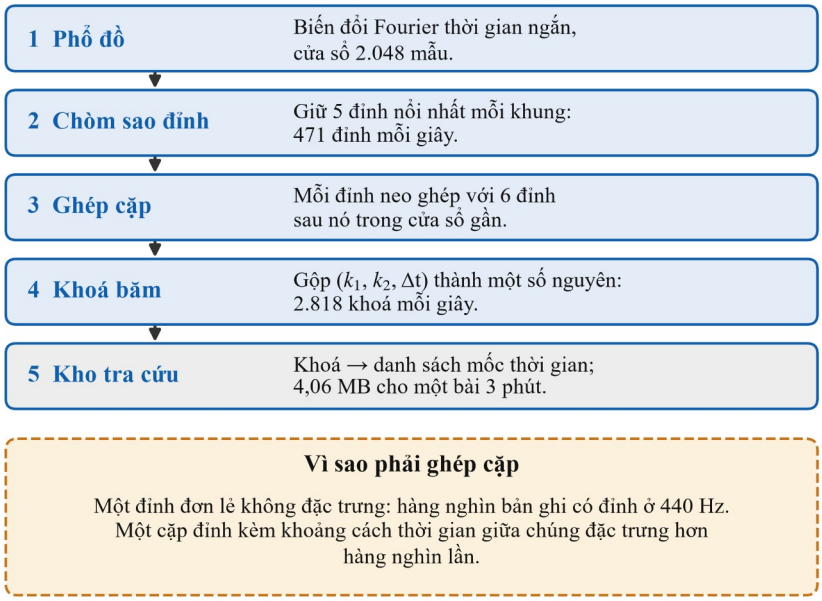
Vân tay giải đúng bài toán còn lại: cho một bản lạ, nó có phải là bản ghi của chúng ta hay không — mà không cần chúng ta đã chuẩn bị gì trong chính bản ghi ấy. Nó không đụng vào tín hiệu; nó chỉ *tóm tắt* tín hiệu thành một thứ tra cứu được.

Bảng 15.13 Ba điều kiện mà một vân tay phải thoả

Điều kiện	Nghĩa là gì
Bền	Không đổi sau nén, sau nhiễu, sau khi qua loa và micrô
Phân biệt được	Hai bản ghi khác nhau phải cho vân tay khác nhau
Tra cứu được	Phải tìm được trong một kho hàng triệu bài trong thời gian chấp nhận được

15.4.2 Chòm sao đỉnh phổ và khoá băm từng cặp

Điều kiện thứ nhất quyết định phải dựng vân tay trên cái gì. Câu trả lời là *đỉnh* phổ — và mục 15.2 đã cho sẵn lý do.

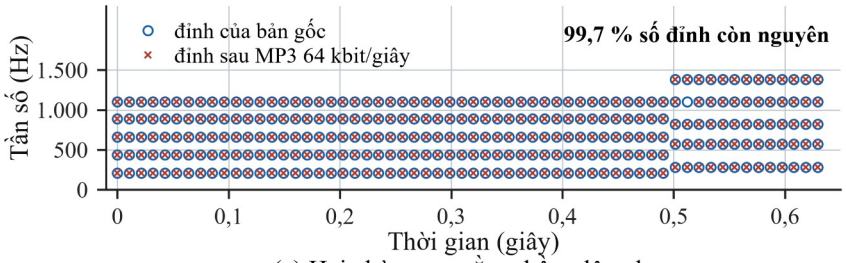


Hình 15.11 Năm bước từ phổ đồ tới một khoá tra cứu

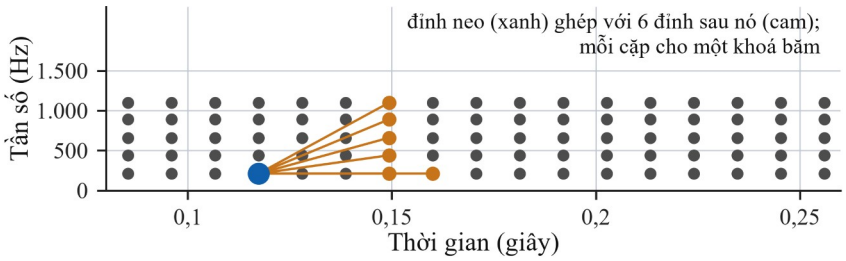
Bảng 15.14 Vì sao đỉnh phổ sống sót

Bản	SI-SDR	Lệch độ lớn phổ (trung vị)	Tỉ lệ đỉnh còn nguyên
MP3 128 kbit/giây	31,2 dB	0,77 dB	99,7 %
MP3 64 kbit/giây	24,9 dB	1,64 dB	99,7 %

Bộ mã hoá loại bỏ những ô nhỏ mà tai không nghe, nhưng nó không dám đụng vào đỉnh — vì đỉnh chính là thứ tai nghe rõ nhất. Đó là toàn bộ lý do phương pháp này chạy được.



(a) Hai chòm sao nằm chồng lên nhau



(b) Ghép cặp quanh một đỉnh neo

Hình 15.12 Hai chòm sao chồng lên nhau, và cách ghép cặp quanh một đỉnh neo

Điều kiện thứ hai — phân biệt được — quyết định cách băm. Một đỉnh đơn lẻ không đặc trưng: hàng nghìn bản ghi có đỉnh ở 440 Hz. Nhưng một *cặp* đỉnh, kèm khoảng cách thời gian giữa chúng, thì đặc trưng hơn hàng nghìn lần.

$$h = (k_1 \ll 20) \text{ OR } (k_2 \ll 8) \text{ OR } \Delta t \tag{15.7}$$

Khoá băm gộp ba số vào một số nguyên: vạch tần số của đỉnh neo, vạch tần số của đỉnh sau nó, và khoảng cách thời gian giữa chúng tính theo số khung.

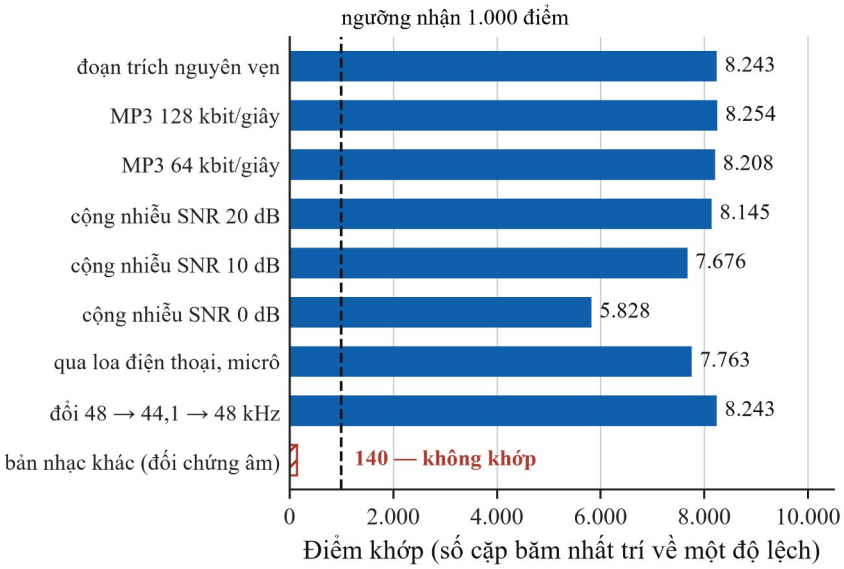
Liệt kê 15.6 Băm từng cặp đỉnh quanh một đỉnh neo

```
def bam_cap_dinh(dinh, quat_t=24, quat_f=48, toi_da=6):
    dinh = sorted(dinh); ra = []
    for i, (t1, f1) in enumerate(dinh):
        dem = 0
        for t2, f2 in dinh[i+1:]:
            dt = t2 - t1
            if dt <= 0: continue
            if dt > quat_t: break
            if abs(f2 - f1) > quat_f: continue
            ra.append((((f1 << 20) | (f2 << 8) | dt), t1))
            dem += 1
            if dem >= toi_da: break
    return ra
```

$$D = \max_{\delta} \left| \{ (h, t) : \; t_{\text{kho}}(h) - t = \delta \} \right| \tag{15.8}$$

Điểm khớp không phải số khoá trùng nhau mà là số cặp khoá cùng trỏ về *một* độ lệch thời gian. Đó là chỗ mạnh của phương pháp: một đoạn trích ba giây vẫn khớp được, vì tất cả các cặp của nó đều nhất trí về cùng một độ lệch, còn những khoá trùng ngẫu nhiên thì rải đều.

15.4.3 Nhận dạng dưới tám phép thử



Hình 15.13 Nhận dạng một đoạn trích ba giây dưới tám phép thử

Bảng 15.15 Truy vấn bằng một đoạn trích ba giây lấy từ giây thứ 2,5

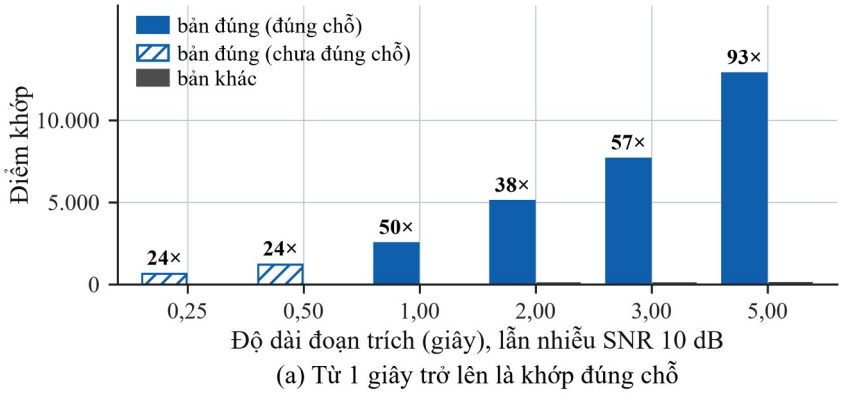
Phép thử	Điểm khớp	Độ lệch trả về	Đúng chỗ
đoạn trích nguyên vẹn	8.243	234	có
MP3 128 kbit/giây	8.254	234	có
MP3 64 kbit/giây	8.208	234	có
cộng nhiễu SNR 20 dB	8.145	234	có
cộng nhiễu SNR 10 dB	7.676	234	có
cộng nhiễu SNR 0 dB	5.828	234	có
qua loa điện thoại rồi micrô	7.763	234	có
đổi 48 → 44,1 → 48 kHz	8.243	234	có
Bản nhạc khác — đối chứng âm	140	469	*Không*

Hai con số đáng nhớ

SNR 0 decibel nghĩa là nhiễu *mạnh bằng* nhạc. Ở mức ấy con người vẫn nhận ra bài hát, và phép khớp vẫn cho 5.828 điểm, đúng chỗ.

Bản nhạc *khác* cho 140 điểm. Khoảng cách giữa 5.828 và 140 là 42 lần — đủ rộng để đặt một ngưỡng mà không phải cân nhắc nhiều. Bài dùng ngưỡng 1.000 điểm.

15.4.4 Đoạn trích ngắn tới đâu, và kho lớn tới đâu



(a) Từ 1 giây trở lên là khớp đúng chỗ

Đỉnh giữ lại mỗi giây	**471**
Khoá băm mỗi giây	**2.818**
Thời gian dựng mỗi giây nhạc	**8,0 ms**
Dung lượng một bài ba phút	**4,06 MB**
Kho cho một triệu bài	**khoảng 4,1 TB**

(b) Giá của một kho vân tay

Hình 15.14 Độ dài đoạn trích cần thiết, và giá của một kho vân tay

Bảng 15.16 Đoạn trích lẫn nhiễu SNR 10 dB, theo độ dài

Độ dài	Điểm với bản đúng	Điểm với bản khác	Tỉ số	Khớp đúng chỗ
0,25 giây	646	27	24×	không
0,50 giây	1.214	51	24×	không
1,00 giây	2.517	50	50×	có
2,00 giây	5.089	133	38×	có
3,00 giây	7.676	134	57×	có
5,00 giây	12.870	139	93×	có

Dưới một giây thì điểm vẫn cao hơn bản khác rất nhiều, nhưng độ lệch thời gian trả về không còn đúng — nghĩa là hệ thống biết "bài này" nhưng không biết "chỗ nào trong bài". Với việc nhận dạng thì thế là đủ; với việc đối chiếu vi phạm thì chưa.

Bảng 15.17 Giá của một kho vân tay

Đại lượng	Giá trị
Đỉnh giữ lại mỗi giây	471
Khoá băm mỗi giây	2.818
Thời gian dựng mỗi giây nhạc	8,0 mili giây
Dung lượng một bài ba phút	4,06 MB
Kho cho một triệu bài	khoảng 4,1 *tb*

Bốn terabyte là một máy chủ, không phải một trung tâm dữ liệu

Đây là lý do nhận dạng bản ghi trở thành dịch vụ phổ thông chứ không còn là công nghệ đắt tiền.

Và cũng là lý do nên dựng kho vân tay cho sản phẩm của chính người đọc: chi phí gần như bằng không, mà nó là cách duy nhất tìm được bản sao khi đã lỡ phát hành bản không có thuỷ vân.

Con số 8 mili giây mỗi giây nhạc nghĩa là dựng kho cho toàn bộ kho nhạc của một đài phát thanh mất vài giờ máy, một lần duy nhất.

15.5 Phát hiện giọng tổng hợp

15.5.1 Bài toán, và cách đặt nó cho đúng

Câu hỏi "giọng này là người hay máy" nghe thì đơn giản, nhưng đặt sai thì không trả lời được. Không có đặc trưng nào của "AI" nói chung; chỉ có đặc trưng của một *cách dựng sóng* cụ thể.

Ba cách đặt bài toán, hai cách sai

Sai: "tìm một đặc trưng phân biệt giọng AI với giọng người". Không có đặc trưng nào như thế, vì AI không phải một thứ mà là rất nhiều kiến trúc khác nhau.

Sai: "huấn luyện một mô hình trên dữ liệu có nhãn". Làm được, nhưng nó chỉ học được những hệ thống có trong dữ liệu, và sai ngay khi gặp hệ thống mới.

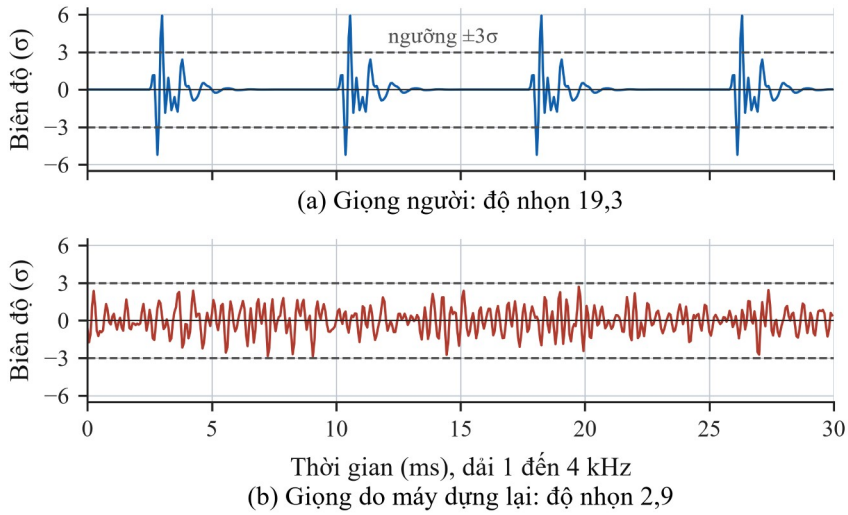
Đúng: "tìm một *dấu vết vật lý* mà một lớp phương pháp cụ thể buộc phải để lại, rồi đo nó". Mục này làm theo cách thứ ba, và cũng đo luôn chỗ cách ấy hết tác dụng.

Lớp phương pháp được chọn là những bộ dựng sóng phải *đoán pha* từ độ lớn phổ — đúng bài toán đã nêu ở mục 13.6. Griffin-Lim là đại diện đơn giản nhất của lớp ấy.

15.5.2 Dấu vết: dạng sóng của giọng người rất nhọn

Nguồn của giọng người là một chuỗi xung thanh môn: dây thanh mở và đóng, mỗi lần đóng phát ra một xung. Mọi hài của giọng đều khoá pha với những xung ấy, nên dạng sóng trong dải trung có những gai rõ rệt.

Bộ dựng sóng đoán pha làm sai đúng sự khoá pha ấy. Các hài vẫn đủ, độ lớn phổ vẫn gần đúng, nhưng chúng không còn cộng lại thành gai nữa — dạng sóng trở nên giống nhiễu Gauss.



(a) Giọng người: độ nhọn 19,3

(b) Giọng do máy dựng lại: độ nhọn 2,9

Hình 15.15 Cùng một câu nói, cùng một phổ mel — khác nhau ở PHA

$$\kappa = \frac{\mathbb{E}[(v-\mu)^4]}{\sigma^4} \tag{15.9}$$

Độ nhọn là mô-men bậc bốn chuẩn hoá. Với phân phối Gauss nó bằng đúng 3. Với một chuỗi xung thì nó lớn hơn nhiều, vì mô-men bậc bốn phạt rất nặng những giá trị xa trung bình.

Liệt kê 15.7 Đo độ nhọn theo từng khung, trong dải 1 đến 4 kHz

```
def do_nhon_theo_khung(x, sr, f1=1000.0, f2=4000.0, n=2048,
                       hop=1024, nguong_rms=0.02):
    xb = loc_dai(x, f1, f2, sr)
    ra = []
    for i in range(0, x.size - n, hop):
        if np.sqrt(np.mean(x[i:i+n]**2)) < nguong_rms:
            continue                          # bo qua doan lang
        v = xb[i:i+n] - xb[i:i+n].mean()
        ra.append(float(np.mean((v / np.std(v))**4)))
    return np.array(ra)
```

Bảng 15.18 Bốn đặc trưng đã thử, trên cùng một cặp giọng

Đặc trưng	Giọng người	Giọng máy	AUC
Độ nhọn dải 1–4 kHz	18,79 ± 5,74	3,44 ± 1,56	1,000
Đỉnh trên hiệu dụng (dB)	16,19 ± 1,05	11,03 ± 1,45	0,981
Tỉ số hài trên nhiễu (dB)	6,07 ± 3,04	1,60 ± 2,55	0,876
Độ phẳng phổ 4–8 kHz	0,490 ± 0,130	0,484 ± 0,077	0,460

$$\mathrm{HNR} = 10\log_{10}\frac{r(T_0)}{1-r(T_0)} \tag{15.10}$$

$$\mathrm{AUC} = \Pr\left(\kappa_{\text{thật}} > \kappa_{\text{máy}}\right) \tag{15.11}$$

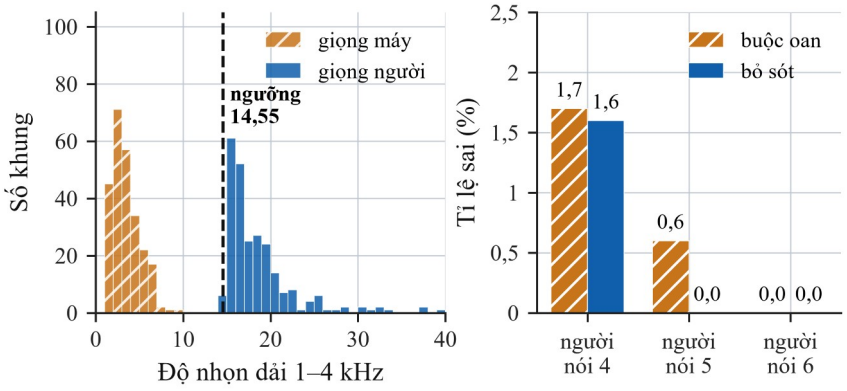
AUC là xác suất một khung giọng người lấy ngẫu nhiên có đặc trưng lớn hơn một khung giọng máy lấy ngẫu nhiên. Giá trị 0,5 nghĩa là đoán ngẫu nhiên; 1,0 nghĩa là hai đống tách hẳn nhau.

Dòng cuối bảng đáng chú ý: độ phẳng phổ vùng cao — đặc trưng mà nhiều người thử đầu tiên — cho AUC 0,460, tức là vô dụng. Nó không vô dụng vì sai nguyên lý mà vì Griffin-Lim giữ *đúng* độ lớn phổ; nó chỉ làm sai pha.

15.5.3 Ngưỡng, tỉ lệ lỗi, và người nói chưa gặp

$$\text{EER}: \quad P_{\text{bỏsót}}(\theta) = P_{\text{buộcoan}}(\theta) \tag{15.12}$$

Khi chưa biết giá của hai loại lỗi thì cách báo kết quả chuẩn là EER: chọn ngưỡng sao cho tỉ lệ bỏ sót bằng tỉ lệ buộc oan, rồi báo chính tỉ lệ ấy.



(a) Hai đống tách hẳn, EER 0,00 %

(b) Người nói chưa gặp

Hình 15.16 Phân bố độ nhọn, ngưỡng, và phép thử trên người nói chưa gặp

Bảng 15.19 Ngưỡng đặt trên một người nói, đem dùng cho ba người khác

	Giọng người (trung bình)	Giọng máy (trung bình)	Buộc oan	Bỏ sót
người nói hiệu chỉnh	18,79	3,44	0,0 %	0,0 %
người nói 4	18,65	3,56	1,7 %	1,6 %
người nói 5	18,69	3,50	0,6 %	0,0 %
người nói 6	18,67	3,49	0,0 %	0,0 %

Điều kiện tối thiểu của một máy dò

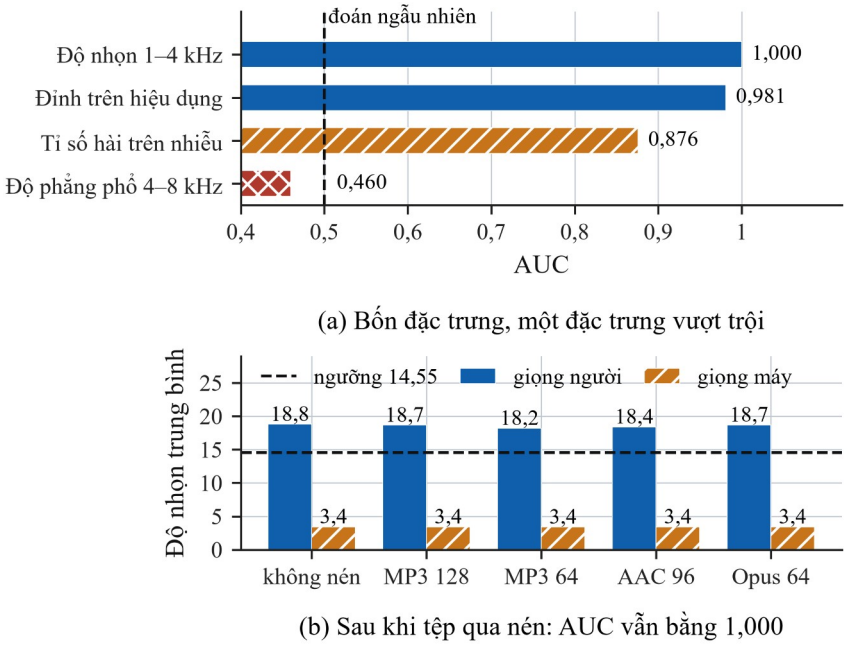
Ngưỡng 14,55 đặt trên một người nói duy nhất, rồi đem dùng cho ba người nói mà máy dò chưa từng thấy: sai nhiều nhất 1,7 phần trăm.

Đó là điều kiện tối thiểu. Một máy dò chỉ đúng trên chính dữ liệu hiệu chỉnh thì không phải máy dò — đây là đúng bài học của mục 13.3.4 về việc đo trên dữ liệu đã học.

Cần nói rõ giới hạn của phép thử này: ba người nói ấy do cùng một mô-đun sinh ra, chỉ khác hạt ngẫu nhiên. Thử trên giọng thu thật là nội dung của bài tập 5.

15.5.4 Máy dò có sống qua nén không

Gần như mọi tệp đến tay người dùng đều đã qua nén ít nhất một lần. Một máy dò chỉ chạy đúng trên tệp không nén thì không dùng được vào việc gì.



(a) Bốn đặc trưng, một đặc trưng vượt trội

(b) Sau khi tệp qua nén: AUC vẫn bằng 1,000

Hình 15.17 Bốn đặc trưng, và máy dò sau khi tệp đã qua nén

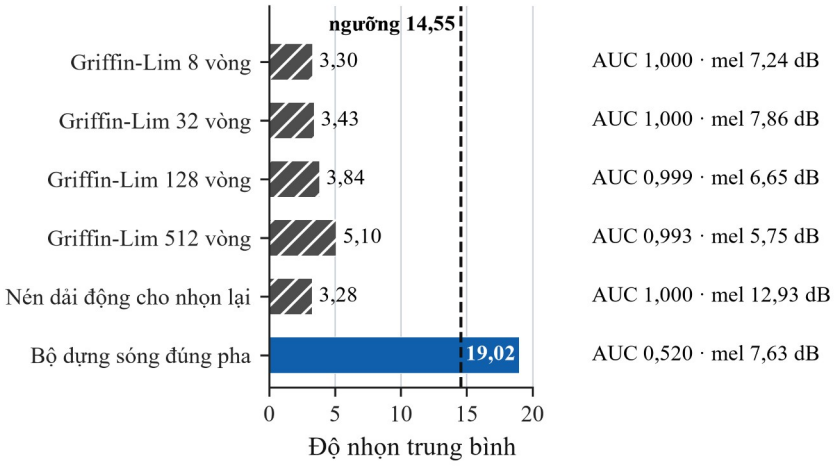
Bảng 15.20 Độ nhọn sau bốn đường nén

Bản	Giọng người	Giọng máy	AUC
không nén	18,79	3,44	1,000
MP3 128 kbit/giây	18,66	3,43	1,000
MP3 64 kbit/giây	18,19	3,40	1,000
AAC 96 kbit/giây	18,40	3,43	1,000
Opus 64 kbit/giây	18,69	3,42	1,000

Nén không xoá được dấu vết, và lý do là ở mục 15.4.2: bộ mã hoá giữ nguyên những đỉnh mà tai nghe rõ, mà chính những đỉnh ấy tạo nên các gai của dạng sóng. Bộ mã hoá không có lý do gì để làm phẳng chúng.

15.5.5 Kẻ làm giả cố chữa — và chỗ máy dò này mù hẳn

Câu hỏi cuối của mục là câu quan trọng nhất: nếu người làm giả biết máy dò đang nhìn vào cái gì thì họ chữa được không.



Hình 15.18 Sáu cách chữa, năm cách không qua được và một cách qua ngay

Bảng 15.21 Sáu cách chữa và kết quả

Cách chữa	Độ nhọn	AUC	Khoảng cách mel	Qua được ngưỡng
chạy Griffin-Lim 8 vòng	3,30	1,000	7,24 dB	không
chạy Griffin-Lim 32 vòng	3,43	1,000	7,86 dB	không
chạy Griffin-Lim 128 vòng	3,84	0,999	6,65 dB	không
chạy Griffin-Lim 512 vòng	5,10	0,993	5,75 dB	không
nén dải động cho nhọn lại	3,28	1,000	12,93 dB	không
bộ dựng sóng dựng *đúng pha*	19,02	0,520	7,63 dB	*Qua*

Dòng cuối là điều quan trọng nhất của cả chương học

Năm cách đầu đều là cố gắng chữa *triệu chứng*, và cả năm đều không qua. Chạy thêm vòng Griffin-Lim có giúp một chút — độ nhọn lên 5,10 sau 512 vòng — nhưng còn cách ngưỡng 14,55 rất xa, và cái giá là gấp mười sáu lần khối lượng tính.

Cách cuối không chữa triệu chứng mà bỏ hẳn nguyên nhân: một bộ dựng sóng dựng *đúng* pha. Độ nhọn của nó là 19,02 — cao hơn cả giọng người — và AUC tụt về 0,520, tức là máy dò đoán ngẫu nhiên.

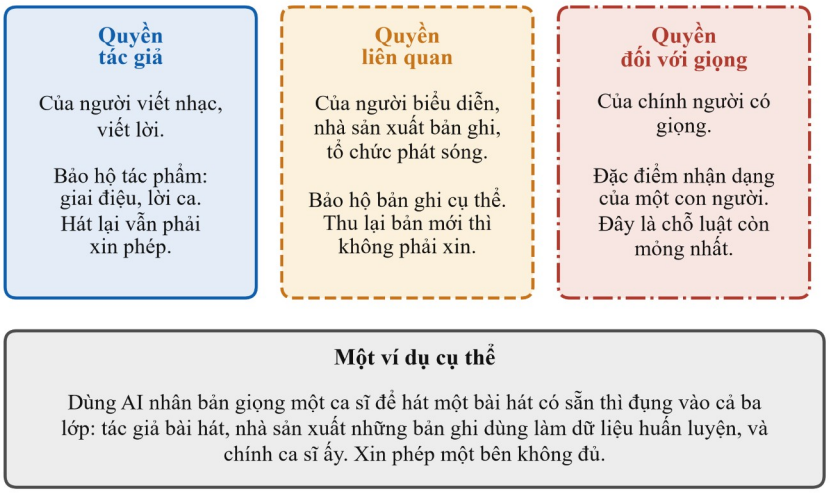
Chú ý cột khoảng cách mel: 7,63 decibel, không tốt hơn Griffin-Lim chút nào. Nghĩa là bản ấy vẫn *sai* y như cũ về độ lớn phổ. Máy dò này không đo chất lượng và không đo "AI"; nó đo *pha*.

Các bộ dựng sóng học được hiện đại chính là loại dựng đúng pha. Máy dò của bài mù hoàn toàn trước chúng — và phải đi tìm dấu vết khác. Đây là một cuộc chạy đua, không phải một bài toán đã giải xong.

15.6 Quyền, đạo đức, và chỗ kỹ thuật dừng lại

15.6.1 Một bản ghi mang ba loại quyền của ba người khác nhau

Sai lầm hay gặp nhất khi nói về bản quyền âm thanh là coi một bản ghi như một vật có một chủ. Thật ra nó mang ít nhất ba lớp quyền chồng lên nhau, và ba lớp ấy thuộc về ba người khác nhau.



Hình 15.19 Ba loại quyền trên cùng một bản ghi

Bảng 15.22 Ba lớp quyền và điều mỗi lớp bảo hộ

Lớp	Của ai	Bảo hộ cái gì	Hát hoặc thu lại thì sao
Quyền tác giả	Người viết nhạc, viết lời	Tác *phẩm*: giai điệu, lời ca	Vẫn phải xin phép
Quyền liên quan	Người biểu diễn, nhà sản xuất bản ghi, tổ chức phát sóng	*Bản ghi* cụ thể	Thu lại bản mới thì không phải xin
Quyền đối với giọng	Chính người có giọng	Đặc điểm nhận dạng của một con người	Đây là chỗ luật còn mỏng nhất

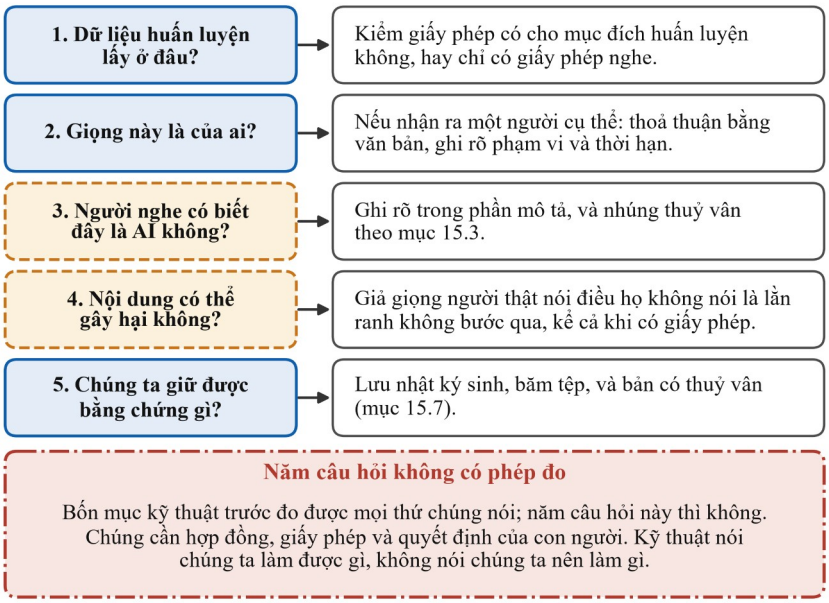
Một ví dụ cụ thể

Dùng AI nhân bản giọng một ca sĩ để hát một bài hát có sẵn thì đụng vào cả ba lớp: tác giả bài hát, nhà sản xuất những bản ghi đã dùng làm dữ liệu huấn luyện, và chính ca sĩ ấy.

Xin phép một bên không đủ. Và ở lớp thứ ba, nhiều nơi trên thế giới còn chưa có quy định rõ ràng để mà xin — đó chính là chỗ đang thay đổi nhanh nhất.

Ở Việt Nam, hai lớp đầu nằm trong Luật Sở hữu trí tuệ, phần quyền tác giả và quyền liên quan. Lớp thứ ba hiện dựa vào quy định về quyền nhân thân đối với hình ảnh và về bảo vệ dữ liệu cá nhân, và còn đang được bàn thêm. Người đọc nên tra văn bản hiện hành thay vì dựa vào trí nhớ, vì đây là phần thay đổi nhanh.

15.6.2 Năm câu hỏi trước khi phát hành



Hình 15.20 Năm câu hỏi phải trả lời trước khi phát hành một bản có giọng do AI sinh

Bảng 15.23 Năm câu hỏi và điều phải làm với mỗi câu

Câu hỏi	Phải làm gì
Dữ liệu huấn luyện lấy ở đâu	Kiểm giấy phép có cho mục đích huấn luyện không — giấy phép nghe không đương nhiên bao gồm việc ấy
Giọng này là của ai	Nếu nhận ra được là một người cụ thể thì phải có thoả thuận bằng văn bản, ghi rõ phạm vi và thời hạn
Người nghe có biết đây là AI không	Ghi rõ trong phần mô tả, và nhúng thuỷ vân theo mục 15.3
Nội dung có thể gây hại không	Giả giọng một người thật nói điều họ không nói là lằn ranh không bước qua, kể cả khi có giấy phép dùng giọng
Ta giữ được bằng chứng gì	Lưu nhật ký sinh, băm tệp, và bản có thuỷ vân — mục 15.7

Chỗ kỹ thuật dừng lại

Bốn mục kỹ thuật trước đo được mọi thứ chúng nói. Năm câu hỏi này thì không: không có phép đo nào trả lời được chúng.

Kỹ thuật nói chúng ta *làm được* gì. Nó không nói chúng ta *nên* làm gì. Ranh giới ấy không phải do luật vẽ ra — luật luôn đi sau — mà do chính người làm nghề tự giữ.

Đây là mục duy nhất trong cả cuốn sách không có một con số nào, và đó là chủ ý.

15.7 Bảo vệ sản phẩm của chính người đọc

15.7.1 Ba công cụ, ba câu hỏi, không cái nào thừa

Ba mục kỹ thuật vừa qua cho ba công cụ. Mục này ghép chúng thành một chuỗi và đo xem chuỗi ấy trả lời được gì.

Phép thử	Thuỷ vân đọc đúng	Sau khi đồng bộ	Điểm vân tay	SHA-256
không đụng gì	100,0 %	100,0 %	22.030	khớp
đổi 1 bit cuối của 1 mẫu	100,0 %	100,0 %	22.030	sai
giảm mức 6 dB	100,0 %	100,0 %	22.030	sai
MP3 128 kbit/giây	100,0 %	100,0 %	22.075	sai
MP3 64 kbit/giây	100,0 %	100,0 %	22.010	sai
đổi 48 → 44,1 → 48 kHz	100,0 %	100,0 %	22.030	sai
trích 3 giây ở giữa	60,9 %	100,0 %	8.091	sai

SHA-256 chỉ trả lời được câu "còn nguyên vẹn từng bit không"; với âm thanh, câu ấy gần như luôn là không.

Ba công cụ không thay được nhau: chúng trả lời ba câu hỏi khác nhau, và một bản giao nghiêm túc cần cả ba.

Hình 15.21 Ba công cụ dưới bảy phép thử

Bảng 15.24 Ba công cụ trên cùng một bản nhạc đã nhúng thuỷ vân

Phép thử	Thuỷ vân đọc đúng	Sau khi đồng bộ lại	Điểm vân tay	SHA-256
không đụng gì	100,0 %	100,0 %	22.030	khớp
đổi 1 bit cuối của 1 mẫu	100,0 %	100,0 %	22.030	*Hỏng*
giảm mức 6 dB	100,0 %	100,0 %	22.030	*Hỏng*
MP3 128 kbit/giây	100,0 %	100,0 %	22.075	*Hỏng*
MP3 64 kbit/giây	100,0 %	100,0 %	22.010	*Hỏng*
đổi 48 → 44,1 → 48 kHz	100,0 %	100,0 %	22.030	*Hỏng*
trích 3 giây ở giữa	60,9 %	100,0 %	8.091	*Hỏng*

Ba công cụ không thay được nhau

SHA-256 không còn khớp ở mọi dòng trừ dòng đầu, kể cả khi chỉ đổi một bit cuối của một mẫu — thứ không ai nghe ra. Nó chỉ trả lời được câu "còn nguyên vẹn từng bit không", và với âm thanh thì câu ấy gần như luôn là *không*.

Nhưng đừng vì thế mà bỏ nó. Nó là công cụ duy nhất chứng minh được rằng một tệp *đúng* là tệp chúng ta đã giao, và đó là điều cần khi có tranh chấp.

Thuỷ vân trả lời "bản này cấp cho ai" — và đọc đúng 100 phần trăm qua mọi phép thử, kể cả trên một đoạn trích ba giây sau khi đồng bộ lại.

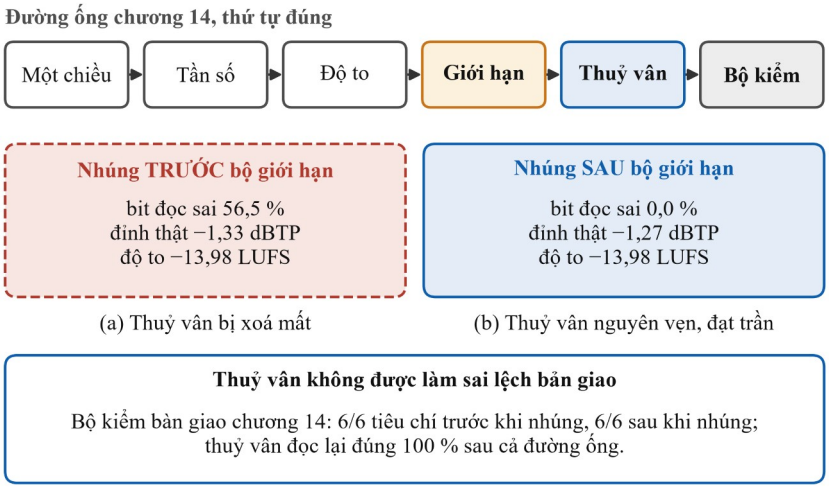
Vân tay trả lời "bản lạ kia có phải của chúng ta" — và nó là công cụ *duy nhất* còn dùng được khi bản nhạc đã phát hành mà bỏ sót nhúng thuỷ vân.

$$N_{\text{vachạm}} \approx 1{,}177\sqrt{2^{b}} \tag{15.13}$$

Về SHA-256: xác suất hai tệp khác nhau cho cùng một mã băm chỉ trở nên đáng kể sau khoảng 2 mũ 128 tệp. Con số ấy lớn hơn mọi thứ có thể đếm được, nên trên thực tế mã băm là định danh tuyệt đối.

15.7.2 Thứ tự: thuỷ vân đặt ở đâu trong đường ống chương 14

Chương 14 dựng một đường ống năm khối và chứng minh rằng thứ tự quyết định kết quả. Thêm thuỷ vân vào đường ống ấy đặt ra đúng câu hỏi cũ: đặt nó ở đâu.



Hình 15.22 Thuỷ vân trước hay sau bộ giới hạn, và bộ kiểm bàn giao

Bảng 15.25 Hai vị trí, hai kết quả

Vị trí	Bit đọc sai	Đỉnh thật	Độ to
nhúng *trước* bộ giới hạn	56,5 %	−1,33 dBTP	−13,98 LUFS
nhúng *sau* bộ giới hạn	0,0 %	−1,27 dBTP	−13,98 LUFS

Vì sao nhúng trước bộ giới hạn thì sai

Bộ giới hạn hạ gain theo thời gian, và lượng hạ khác nhau ở mỗi chỗ. Thuỷ vân trải phổ đọc bằng tương quan, mà tương quan thì bị phá khi từng đoạn của chuỗi bị nhân với những hệ số khác nhau.

Nhúng *sau* bộ giới hạn thì thuỷ vân nguyên vẹn — nhưng phải kiểm lại đỉnh thật, vì thuỷ vân cộng thêm năng lượng. Ở đây đỉnh thật đi từ −1,33 lên −1,27 dBTP, vẫn dưới trần −1,0 nên đạt.

Nếu thuỷ vân đẩy đỉnh thật vượt trần thì cách chữa là hạ trần bộ giới hạn xuống −1,5 dBTP trước, chứ không phải hạ mức thuỷ vân — hạ mức thuỷ vân là mất độ bền, theo đúng bảng ở mục 15.3.7.

Bảng 15.26 Bộ kiểm bàn giao của chương 14, trước và sau khi nhúng

	Trước khi nhúng	Sau khi nhúng
Số tiêu chí đạt	6 / 6	6 / 6
Thuỷ vân đọc lại	—	100,0 %

Đây là điều kiện bắt buộc: bảo vệ *không* được làm sai lệch bản giao. Một thuỷ vân làm tệp trượt bộ kiểm là một thuỷ vân đặt sai chỗ hoặc đặt quá mạnh.

15.7.3 Giá của cả chuỗi

$$C_{\text{bảovệ}} = \sum_{i} c_i \quad [\text{mschomỗigiây}] \tag{15.14}$$

Bảng 15.27 Thời gian máy cho mỗi giây âm thanh

Công cụ	Thời gian	Trả lời câu hỏi
Nhúng thuỷ vân	1,7 ms	Bản này chúng ta cấp cho ai
Dựng kho vân tay	9,3 ms	Bản lạ kia có phải của chúng ta
Băm SHA-256	0,5 ms	Tệp có còn nguyên vẹn từng bit
Tổng	11,5 ms	—

Mười một mili giây cho mỗi giây âm thanh

Nghĩa là bảo vệ một bản ba phút tốn khoảng hai giây máy. So với 85 mili giây mỗi giây của cả đường ống hậu kỳ ở mục 14.6.3 thì đây là phần rẻ nhất.

Không có lý do kỹ thuật nào để bỏ qua bước này. Lý do duy nhất người ta bỏ qua là bỏ sót — và bỏ sót thì không sửa được về sau, ít nhất là với thuỷ vân.

Vì thế nên đưa cả ba vào chính đường ống, ngay sau bộ giới hạn và ngay trước bộ kiểm, để chúng chạy tự động chứ không phụ thuộc trí nhớ của ai.

15.8 Dự án thực hành — giai đoạn 5: hoàn thiện và trình bày

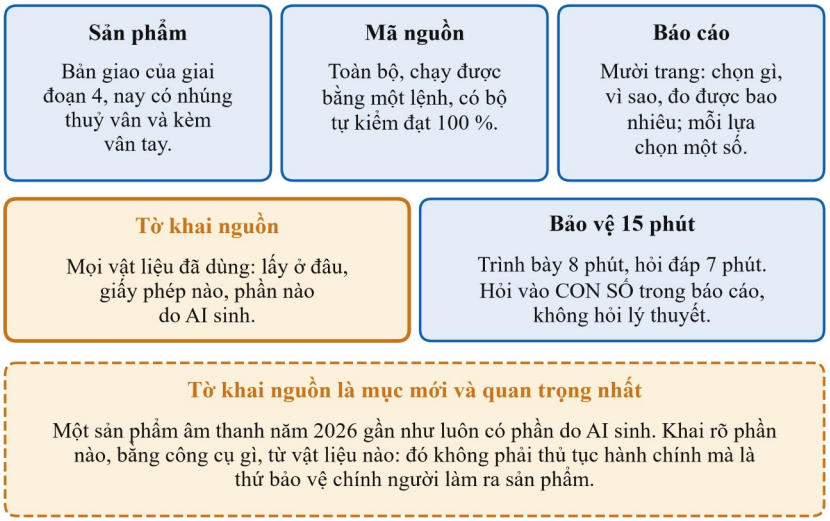
15.8.1 Giai đoạn 5 gộp lại toàn bộ bốn giai đoạn trước

Bảng 15.28 Năm giai đoạn của dự án thực hành

Giai đoạn	Mốc	Chương	Trạng thái
1	Sinh và đo tín hiệu	Chương 4	Làm sau chương 4
2	Chuỗi hiệu ứng có đo	Chương 9	Làm sau chương 9
3	Bản trộn chạy trên trình duyệt	Chương 12	Làm sau chương 12
4	Bản giao đạt bộ kiểm	Chương 14	Làm sau chương 14
5	Lớp bảo vệ, tài liệu và trình bày hoàn chỉnh	Chương 15	Làm sau chương này

Giai đoạn 5 không yêu cầu làm lại gì. Nó yêu cầu hai việc: bọc sản phẩm của giai đoạn 4 bằng lớp bảo vệ của mục 15.7, và trình bày được mọi lựa chọn đã làm trong suốt bốn giai đoạn — bằng số đo, không bằng cảm nhận.

15.8.2 Năm kết quả cần có



Hình 15.23 Năm kết quả cần có ở giai đoạn 5

Bảng 15.29 Năm kết quả cần có và tiêu chuẩn của từng thứ

Kết quả	Tiêu chuẩn
Sản phẩm	Bản giao của giai đoạn 4, nay có nhúng thuỷ vân đọc lại được, kèm tệp kho vân tay dựng từ chính nó và mã băm SHA-256 của tệp giao
Mã nguồn	Toàn bộ, chạy được bằng một lệnh từ vật liệu nguồn tới tệp giao; có bộ tự kiểm riêng của dự án và bộ ấy phải đạt 100 %
Báo cáo	Mười trang. Mỗi lựa chọn tham số phải dẫn ít nhất một con số đo được; mỗi chỗ *không* đo được phải nói rõ là không đo được
Tờ khai nguồn	Mọi vật liệu đã dùng: lấy ở đâu, giấy phép nào, phần nào do AI sinh và bằng công cụ gì
Trình bày 15 phút	Trình bày 8 phút, hỏi đáp 7 phút với người nghe. Câu hỏi nên hỏi vào *con số* trong báo cáo, không hỏi lý thuyết chung

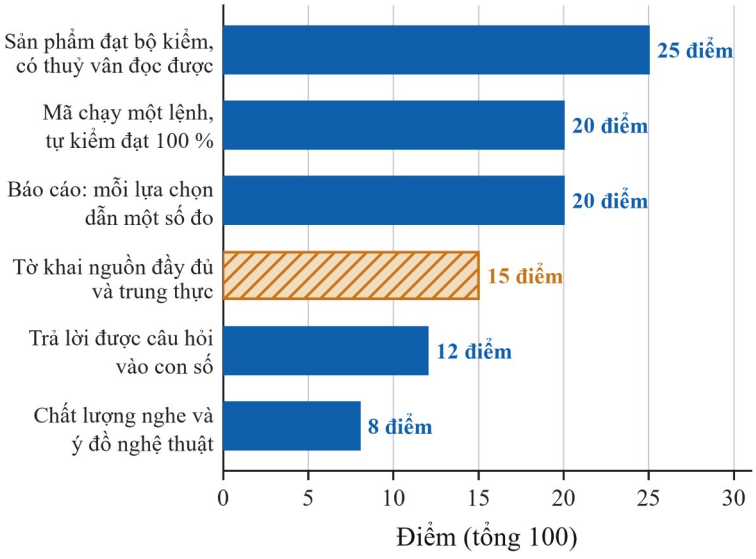
Tờ khai nguồn là mục mới, và là mục quan trọng nhất

Một sản phẩm âm thanh làm năm 2026 gần như luôn có phần do AI sinh: một đoạn nền, một giọng đọc, một mẫu trống. Khai rõ phần nào, bằng công cụ gì, từ vật liệu nào — đó không phải thủ tục hành chính.

Nó là thứ bảo vệ chính người làm ra sản phẩm. Khi có tranh chấp, người có tờ khai đầy đủ ở thế hoàn toàn khác người không có.

Và nó là chỗ duy nhất trong cả cuốn sách mà sự trung thực không kiểm được bằng máy. Vì thế càng phải khai cho nghiêm: không bỏ sót một vật liệu nào.

15.8.3 Tự đánh giá kết quả



Hình 15.24 Các hạng mục tự đánh giá giai đoạn 5

Khi tự đánh giá giai đoạn này, hãy kiểm những thứ không dựa vào tai trước: sản phẩm đạt bộ kiểm chương 14 và có thuỷ vân đọc được; mã chạy một lệnh và tự kiểm đạt 100 phần trăm; báo cáo mà mỗi lựa chọn dẫn một số đo; tờ khai nguồn đầy đủ và trung thực; và trả lời được câu hỏi hỏi thẳng vào một con số. Phần nghe để sau cùng. Đó không phải vì tai không quan trọng, mà vì phần nghe chỉ có ý nghĩa khi phần kỹ thuật đã sạch.

15.8.4 Bảy câu hỏi nên tự trả lời được khi trình bày dự án

Danh sách dưới đây nên xem từ trước, vì mục đích của nó không phải làm khó mà là kiểm xem người làm có thật sự làm những phép đo đã viết trong báo cáo hay không.

- Tần số lấy mẫu của mô hình nguồn là bao nhiêu, và bạn biết điều đó bằng cách nào?
- Số cạnh của bộ lọc đổi tần số là bao nhiêu, và vì sao lại là con số ấy chứ không phải gấp đôi hay một nửa?
- Hệ số tương quan giữa hai đoạn tại mối nối dài nhất là bao nhiêu, và bạn chọn luật mờ chéo nào theo con số ấy?
- Đỉnh thật của bản giao là bao nhiêu, và bạn đo nó ở bước nào trong đường ống?
- Mức nhúng thuỷ vân là bao nhiêu decibel, và bạn đã thử tấn công nào để kiểm độ bền?
- Chạy lại mã hai lần có cho ra tệp hoàn toàn giống không, và bạn kiểm điều đó bằng lệnh gì?
- Phần nào của sản phẩm do AI sinh, và bạn dựa vào giấy phép nào để dùng phần ấy?

> **Cách trả lời đúng và cách trả lời sai**
>
> Trả lời *sai*: "tôi chọn c bằng 16 vì đó là giá trị thường dùng". Câu ấy đúng về sự thật nhưng không chứng minh được người làm đã đo gì.
>
> Trả lời *đúng*: "tôi chọn c bằng 16 vì với c bằng 8 thì dải giữ nguyên vẹn chỉ còn 76,7 phần trăm Nyquist, còn với c bằng 32 thì độ trễ lên 99,5 mili giây mà độ triệt chỉ hơn 1 decibel. Bảng đo ở trang 6 của báo cáo."
>
> Khác nhau không phải ở độ dài câu trả lời mà ở chỗ có dẫn số đo hay không. Mỗi câu hỏi trong bảy câu trên đều có một con số cụ thể làm câu trả lời, và con số ấy phải nằm trong báo cáo.

Bảng 15.30 Ba lỗi hay gặp nhất khi trình bày dự án

Lỗi	Vì sao nó xảy ra	Cách tránh
Không nhớ con số của chính mình	Báo cáo do một người viết, người khác lên trình bày	Mọi người tham gia cùng chạy lại mã một lần trước khi trình bày
Trả lời bằng lý thuyết chung	Đọc tài liệu nhưng không tự đo	Mỗi khẳng định trong báo cáo phải có một dòng lệnh sinh ra nó
Không giải thích được chỗ mình *không* đo được	Ngại nói ra giới hạn của mình	Nói rõ giới hạn là điểm *cộng*, không phải điểm trừ

15.9 Tổng kết chương và tổng kết cuốn sách

15.9.1 Ba bài thực hành

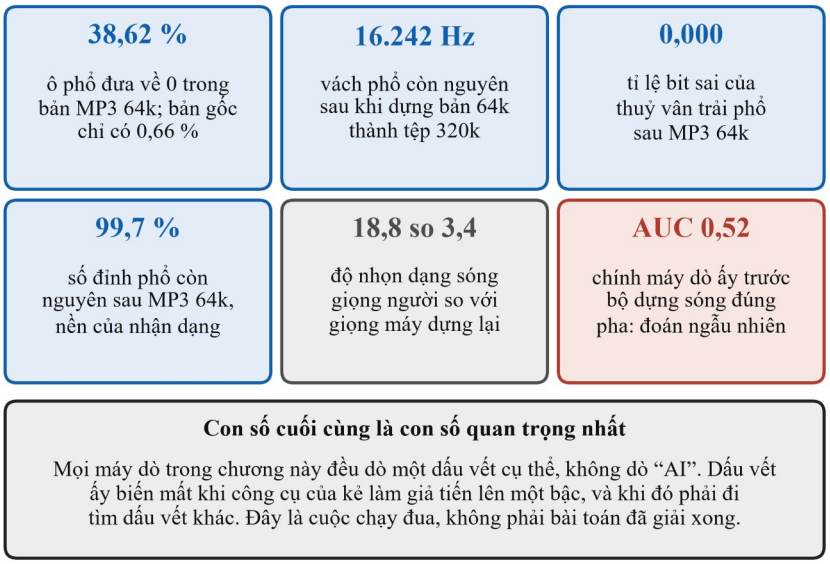
- **Thực hành 1 — dấu vết codec.** Nén một đoạn nhạc bằng ffmpeg ở bốn bitrate, đo vách phổ và tỉ lệ ô phổ rỗng của từng bản, rồi thử đoán bitrate của một tệp do người khác đưa cho.
- **Thực hành 2 — thuỷ vân và tấn công.** Nhúng thuỷ vân trải phổ ở mức −26 decibel, rồi tự nghĩ ra ba phép tấn công không có trong bảng ở mục 15.3.5 và đo tỉ lệ bit sai của từng phép.
- **Thực hành 3 — máy dò giọng máy.** Dựng kho vân tay cho một đoạn nhạc, rồi thử truy vấn bằng chính đoạn ấy phát qua loa điện thoại và thu lại bằng micrô máy tính. So điểm khớp với bảng ở mục 15.4.3.

15.9.2 Bảy bài tập

- **Bài 1 — điều tra một tệp lạ.** Lấy ba tệp nhạc trên máy của người đọc, đo vách phổ, tỉ lệ ô rỗng và tương quan hai kênh, rồi viết một đoạn ngắn kết luận mỗi tệp đã đi qua những khâu nào.
- **Bài 2 — dựng lại phép thử chuyển mã.** Nén một đoạn ở 64 kbit/giây, giải nén, nén lại ở 320, và chứng minh bằng hai con số rằng bản cuối không phải bản chất lượng cao.
- **Bài 3 — thêm mẫu đồng bộ.** Thêm một mẫu đồng bộ vào sơ đồ trải phổ, rồi đo lại hàng "cắt 1.000 mẫu đầu" của bảng ở mục 15.3.5. Nói rõ cái giá phải trả về dung lượng.
- **Bài 4 — vân tay cho kho nhạc của người đọc. Dựng kho vân tay cho hai mươi bài trên máy của mình**, rồi truy vấn bằng đoạn trích ba giây của từng bài và đo tỉ lệ nhận đúng cùng điểm của bản đứng thứ hai.
- **Bài 5 — máy dò trên giọng thật.** Thu giọng của ba người, chạy chúng qua vòng phổ mel và Griffin-Lim, rồi dựng lại bảng ở mục 15.5.3 trên vật liệu ấy. Ngưỡng 14,55 có còn dùng được không?

– **Bài 6 — một đặc trưng thứ năm.** Nghĩ ra một đặc trưng khác để phân biệt giọng người với giọng do máy dựng lại, đo AUC của nó, và giải thích nó đang nhìn vào dấu vết vật lý nào.
– **Bài 7 — tờ khai nguồn.** Viết tờ khai nguồn cho chính dự án thực hành của mình, theo đúng năm câu hỏi ở mục 15.6.2. Đây là bài tập duy nhất không có phép đo nào.

15.9.3 Sáu con số của chương 15

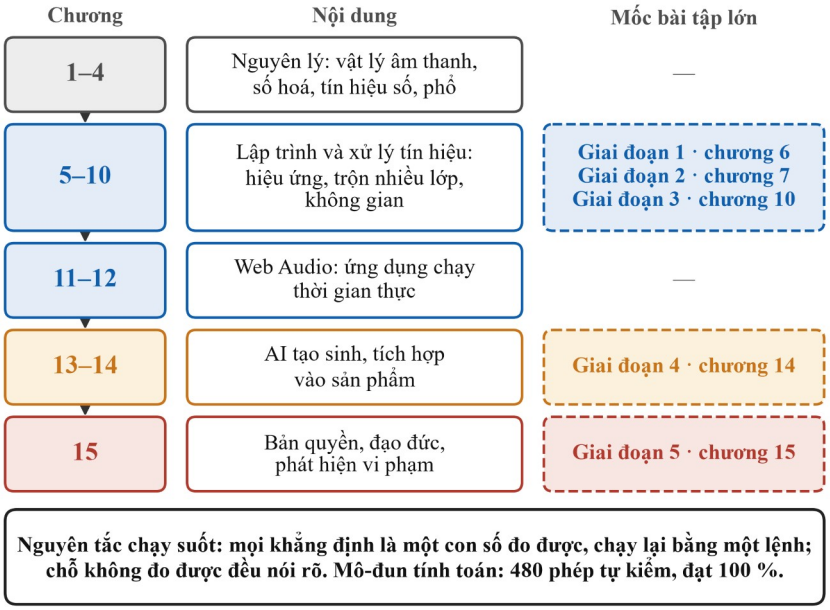


Hình 15.25 Chương 15 gói lại trong sáu con số

Bảng 15.31 Sáu con số và ý nghĩa của từng con số

Con số	Của cái gì	Nói lên điều gì
38,62 %	ô phổ bị đưa về 0 trong một bản MP3 64k	Bản gốc chỉ có 0,66 % — dấu vết nén không giấu được
16.242 Hz	vách phổ sau khi dựng bản 64k thành tệp 320k	Bitrate ghi trong tệp không nói gì về chất lượng thật
0,000	tỉ lệ bit sai của thuỷ vân trải phổ sau MP3 64k	Giấu đúng chỗ codec không dám đụng thì thuỷ vân sống
99,7 %	số đỉnh phổ còn nguyên sau MP3 64k	Nền của mọi phép nhận dạng bản ghi
18,8 so 3,4	độ nhọn dạng sóng của giọng người so với giọng máy	Một đặc trưng, AUC 1,000, sống qua mọi phép nén
AUC 0,52	chính máy dò ấy khi gặp bộ dựng sóng đúng pha	Máy dò dò *dấu vết*, không dò "AI" — và dấu vết thì biến mất

15.9.4 Tổng kết cuốn sách



Hình 15.26 Mười lăm chương trên một đường đi

Mười lăm chương đi từ vật lý của sóng âm tới câu hỏi một bản ghi có được phép tồn tại không. Đường đi ấy dài, nhưng nó có một nguyên tắc chạy suốt.

Một nguyên tắc chạy suốt mười lăm chương

Không có câu nào trong cuốn sách này dựa vào "nghe thì thấy hay". Mọi khẳng định đều là một con số đo được, chạy lại được bằng một lệnh.

Và mỗi chỗ *không* đo được đều được nói rõ là không đo được: hiện tượng lệch phơi nhiễm ở mục 13.4.3, giới hạn của bài toán đồ chơi ở mục 13.5.3, phép dò lưới khung ở mục 15.2.4, và cả mục 15.6 không có một con số nào.

Mô-đun tính toán của cuốn sách khép lại với 480 phép tự kiểm, chạy bằng một lệnh, đạt 100 phần trăm.

Bảng 15.32 Năm thói quen đáng mang ra khỏi cuốn sách

Thói quen	Học ở chương nào
Đo trước, rồi mới kết luận — và đo ở *cuối* đường xử lý	Chương 7 · 14.4.5 · 14.6
Tính trần trên của một bài toán trước khi bắt tay giải nó	13.7 · 15.4.2
Hỏi xem thước đo đang phạt cái gì, trước khi tin một con số kém	13.8 · 14.6.2 · 15.3.4
Dựng phép đo có kiểm soát, lặp lại nhiều lần, rồi lấy trung vị	14.5.3
Nói rõ chỗ mình không đo được, thay vì lặng lẽ bỏ qua	13.5.3 · 15.2.4 · 15.5.5

Thứ đáng mang ra khỏi cuốn sách không phải danh sách công thức — công thức tra lại được bất cứ lúc nào. Thứ đáng mang đi là năm thói quen trong bảng

trên, và chúng dùng được cho bất kỳ lĩnh vực kỹ thuật nào chứ không riêng âm thanh.

15.9.5 Sau cuốn sách này thì đi tiếp theo hướng nào

Cuốn sách dừng ở đây, nhưng mỗi phần đều mở ra một hướng đi riêng. Bảng dưới đây là gợi ý cho người muốn đi tiếp, kèm chỗ trong sách mà hướng ấy bắt đầu.

Bảng 15.33 Năm hướng đi tiếp, và chỗ bắt đầu của mỗi hướng

Hướng	Bắt đầu từ mục nào	Việc làm được sau đó
Xử lý tín hiệu thời gian thực	Chương 11 và chương 12 — Web Audio, ngân sách trễ ở 14.6.3	Viết plugin âm thanh, làm thiết bị nhúng, dựng hệ thống biểu diễn
Học máy cho âm thanh	Chương 13 — token, khuếch tán, tách nguồn	Huấn luyện mô hình sinh hoặc mô hình nhận dạng cho tiếng Việt
Xử lý hậu kỳ và phát hành	Chương 10 và chương 14 — trộn, chuẩn hoá, bộ kiểm bàn giao	Làm kỹ thuật viên hậu kỳ, dựng đường ống phát hành tự động
Pháp y âm thanh và bảo vệ nội dung	Chương 15 — dấu vết codec, thuỷ vân, vân tay	Làm hệ thống giám sát bản quyền, hỗ trợ giám định
Âm học và không gian	Chương 2 và chương 10 — sóng, phòng, âm thanh vòm	Thiết kế phòng thu, xử lý âm học, dựng âm thanh cho phim

Ba việc nên làm ngay trong tháng đầu sau khi đọc xong cuốn sách

Giữ lại mô-đun tính toán và bộ tự kiểm của chính mình, rồi thêm vào đó mỗi khi gặp một phép đo mới. Sau hai năm nó sẽ là tài sản nghề nghiệp đáng giá hơn bất kỳ ghi chép nào.

Chạy bộ kiểm bàn giao của chương 14 trên ba sản phẩm cũ của chính mình. Kết quả thường gây ngạc nhiên, và đó là cách nhanh nhất để thấy cuốn sách này có ích ở đâu.

Dựng kho vân tay cho mọi thứ đã phát hành. Mất vài giờ máy, làm một lần, và nó là cách duy nhất tìm lại được bản sao về sau.

Chúc bạn đọc làm ra những sản phẩm mà chính mình đo được, giải thích được, và đứng tên được.

Tài liệu tham khảo của chương

Nguồn trực tiếp của nội dung trong bài

- Brandenburg, K. — MP3 and AAC Explained. Nguồn của phần cấu trúc khung, stereo ghép đôi và tiền vọng ở mục 15.2.

- Wang, A. — An Industrial-Strength Audio Search Algorithm. Bài gốc của phương pháp chòm sao đỉnh phổ và khoá băm từng cặp; toàn bộ mục 15.4 là một bản cài đặt rút gọn của nó.
- Cox, I. J. và cộng sự — Secure Spread Spectrum Watermarking for Multimedia. Nguồn của sơ đồ trải phổ ở công thức (15.3) và (15.4), kể cả lập luận về chỗ đặt thuỷ vân.
- Bender, W. và cộng sự — Techniques for Data Hiding. Nguồn của sơ đồ ẩn bằng tiếng vọng ở công thức (15.6), và của phép đọc bằng cepstrum.
- Griffin, D., Lim, J. — Signal Estimation from Modified Short-Time Fourier Transform. Đã dẫn ở chương 13; ở chương này nó là bộ dựng sóng dùng làm vật liệu cho mục 15.5.
- Luật Sở hữu trí tuệ của Việt Nam, phần quyền tác giả và quyền liên quan, cùng các văn bản hướng dẫn hiện hành. Nền pháp lý cho hai lớp quyền đầu ở mục 15.6.1.

Sách và tài liệu nền

- Bosi, M., Goldberg, R. E. — Introduction to Digital Audio Coding and Standards. Sách nền đầy đủ nhất cho mọi thứ ở mục 15.2.
- Cox, I. J., Miller, M. L., Bloom, J. A. — Digital Watermarking and Steganography. Sách nền cho mục 15.3, kể cả phần về mất đồng bộ và mẫu đồng bộ.
- Cano, P. và cộng sự — A Review of Audio Fingerprinting. Đối chiếu phương pháp ở mục 15.4 với các họ phương pháp khác.
- Müller, M. — Fundamentals of Music Processing. Chương về nhận dạng và đối sánh nội dung âm nhạc; có mã Python đi kèm.

Công cụ nên biết

- ffmpeg. Công cụ dùng cho mọi phép nén trong bài. Lệnh cần thuộc: ffmpeg -i vao.wav -c:a libmp3lame -b:a 128k ra.mp3.
- Chromaprint và AcoustID — bộ vân tay âm thanh mã nguồn mở, dùng được ngay. Đối chiếu với bản tự viết ở mục 15.4.
- Spek hoặc Audacity với chế độ phổ đồ. Cách nhanh nhất để *nhìn* thấy vách phổ mà không cần viết mã — hữu ích khi cần kiểm nhanh một tệp.
- Bộ tự kiểm của mô-đun đi kèm sách. Ba mươi hàm mới của chương này đều có phép kiểm riêng, và người đọc chạy chính bộ ấy trên mã của mình để tự đối chiếu.

Đọc thêm theo hướng

- **Muốn đi sâu về pháp y âm thanh:** tìm đọc về phân tích tần số điện lưới trong bản ghi — một phương pháp xác định thời điểm thu rất mạnh, dựa trên đúng ý tưởng "dấu vết không cố ý" của mục 15.2.
- **Muốn đi sâu về thuỷ vân:** tìm đọc về thuỷ vân theo miền biến đổi và về tấn công cộng tác — tình huống nhiều bản có thuỷ vân khác nhau bị trộn lại để xoá dấu.

- **Muốn đi sâu về phát hiện giọng tổng hợp:** tìm đọc về các cuộc thi ASVspoof và về khái niệm tấn công phát lại. Chúng cho thấy rõ tính chạy đua mà mục 15.5.5 nói tới.
- **Muốn đi sâu về khung pháp lý:** theo dõi các quy định mới về ghi nhãn nội dung do AI sinh. Đây là phần thay đổi nhanh nhất, và mọi điều viết trong bài đều cần tra lại văn bản hiện hành.

Một lưu ý về các con số trong bài

> **Cách đọc các con số**
>
> Mọi phép nén trong bài chạy bằng ffmpeg với libmp3lame, aac, libopus và libvorbis. Vì thế các con số về vách phổ và tỉ lệ ô rỗng phản ánh hành vi *thật* của các bộ mã hoá ấy, không phải của một mô hình giả lập — nhưng chúng cũng phụ thuộc vào phiên bản ffmpeg và vào tham số mặc định của từng bộ mã hoá.
>
> Các con số về tỉ lệ bit sai, điểm khớp vân tay, độ nhọn và AUC là kết quả đo trên vật liệu sinh bằng chính mô-đun của cuốn sách, nên chạy lại được hoàn toàn như cũ. Đổi vật liệu thì con số đổi theo; *hình dạng* của kết quả thì không.
>
> Giọng dùng ở mục 15.5 là giọng *mô phỏng*, không phải giọng thu thật, và "giọng máy" là kết quả của Griffin-Lim chứ không phải của một mô hình sinh giọng. Giới hạn ấy nêu rõ ở mục 15.1.3, và bài tập 5 yêu cầu chạy lại trên giọng thu thật.
>
> Mục 15.6 không có một con số nào. Đó là mục duy nhất như thế trong cả cuốn sách, và lý do nêu ngay tại mục.
>
> Mã Python in trong bài là mã đã chạy, không phải mã giả.