TRƯỜNG ĐẠI HỌC XÂY DỰNG HÀ NỘI
KHOA CÔNG NGHỆ THÔNG TIN

LẬP TRÌNH ÂM THANH

CHƯƠNG 13

AI tạo sinh nhạc và giọng

Sau chương này người học phải làm được

- So sánh bốn cách biểu diễn âm thanh cho mô hình; tính số lượng và bitrate của mỗi cách.
- Giải thích bằng con số vì sao không mô hình nào sinh thẳng dạng sóng.
- Cài lượng tử hoá dư; đo chất lượng tái dựng theo số tầng và số từ mã.
- Phân biệt phép đo trên dữ liệu đã học với phép đo trên phần giữ riêng.
- Đo entropy chéo của mô hình tự hồi quy; chọn nhiệt độ, top-k, top-p có căn cứ.
- Giải thích quá trình thuận của khuếch tán, lịch cosine, và quan hệ số bước với chất lượng.
- Chỉ ra bài toán pha; tính trần trên của tách nguồn; chọn thước đo và thiết kế phép nghe có đối chứng.

MỤC 13.1

Vị trí chương học trong học phần

- Chương đầu tiên của học phần nói về trí tuệ nhân tạo, nhưng không bắt đầu từ đầu.
- Gần như mọi khái niệm cần dùng đã có từ mười hai chương trước.
- Chương không huấn luyện mô hình lớn; chương đo bằng số các cơ chế quyết định chất lượng.

13.2	Biểu diễn	Sóng, phổ, mel hay token, và giá của mỗi cách.
13.3	Codec nơ-ron	Lượng tử hoá dư: nơi token của nhạc sinh ra.
13.4	Tự hồi quy	Sinh từng token, và cái giá tuần tự của nó.
13.5	Khuếch tán	Sinh cả đoạn một lúc, đổi số bước lấy chất lượng.
13.6	Giọng và bài toán pha	Vì sao phổ mel chưa đủ để dựng lại tiếng.
13.7	Tách nguồn	Mặt nạ, trần trên, và hai cách không cần học.
13.8	Đánh giá	Ba chỗ thước đo khách quan bị đánh lừa.

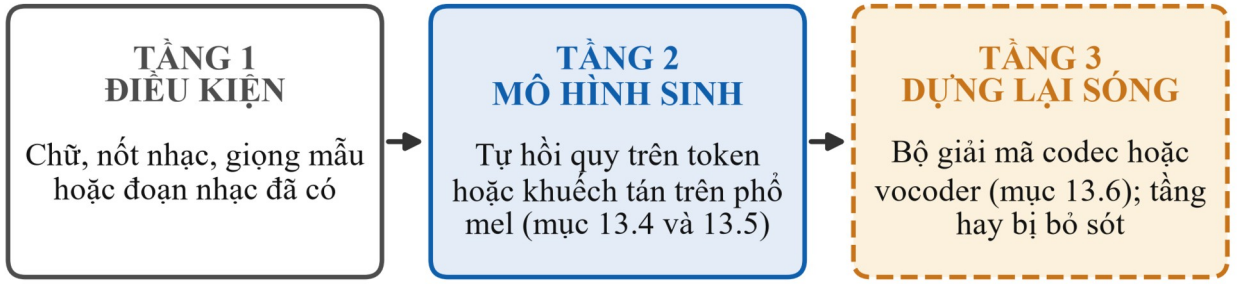
Chương này không huấn luyện mô hình lớn nào: chỉ đo những yếu tố quyết định chất lượng của mọi mô hình như thế, bằng mã chạy trên máy tính xách tay.

Đo được mà không cần mô hình lớn

Lượng tử hoá dư cài được bằng numpy. Quá trình thuận của khuếch tán tính được bằng công thức. Tách nguồn có những trần trên tính được. Chỗ không đo được: lệch phơi nhiễm (13.4.3) và giới hạn của bài toán đồ chơi (13.5.3).

Ba tầng của một hệ thống sinh

- Tầng 1 là điều kiện; tầng 2 là mô hình sinh; tầng 3 là dựng lại sóng.
- Phần lớn hiểu lầm đến từ việc gộp ba tầng làm một.
- Chương xoay quanh cái đi giữa ba tầng: bốn cách biểu diễn.



Cái đi giữa ba tầng ấy quyết định tất cả:

Dạng sóng thô	44.100 số mỗi giây	Không mô hình nào sinh nổi (13.2.1)
Phổ độ lớn	88.372 số mỗi giây	Vẫn quá nhiều, và mất pha (13.6)
Phổ mel	13.781 số mỗi giây	Đích của họ khuếch tán; cần vocoder
Token rời rạc	1.378 số mỗi giây	Đích của họ tự hồi quy (13.3)

Bốn dòng trên cách nhau hai bậc độ lớn; mọi quyết định kiến trúc bắt đầu từ đó.

Hình 13.2 — Ba tầng của một hệ thống sinh, và bốn cách biểu diễn đi giữa chúng

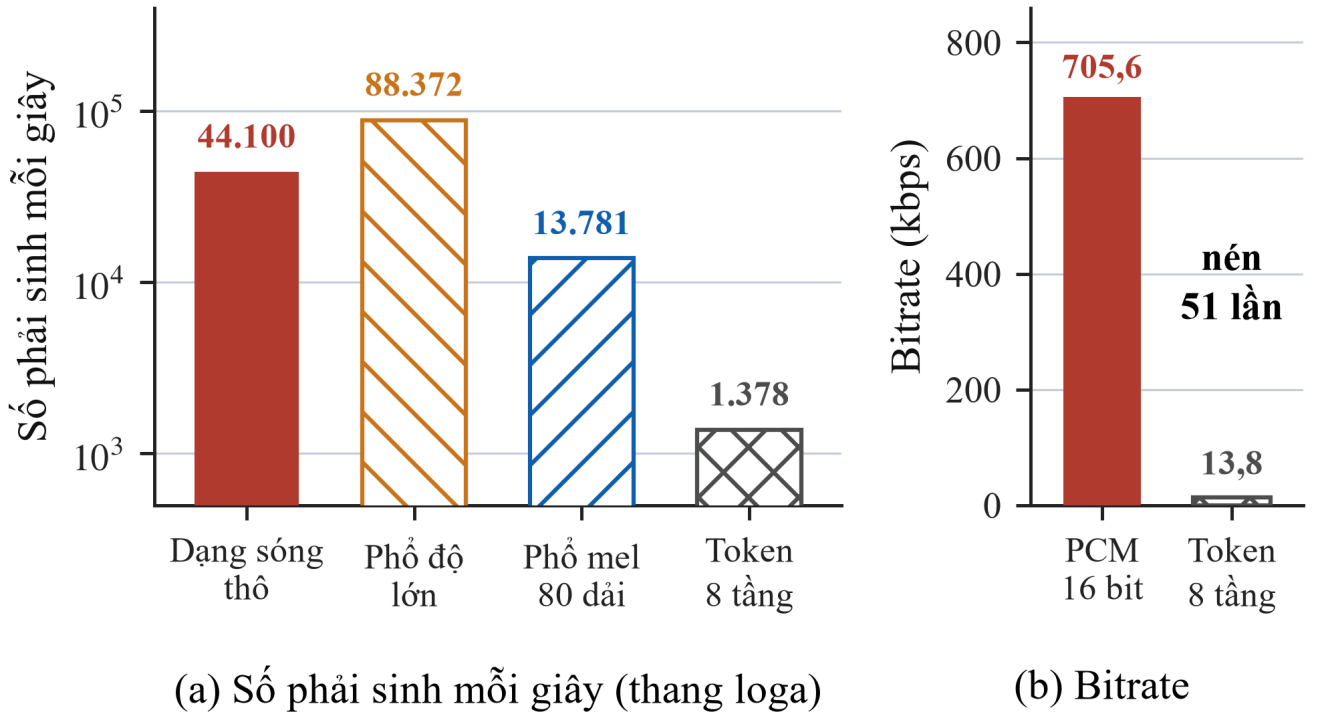
MỤC 13.2

Biểu diễn âm thanh cho một mô hình

- Câu hỏi đầu tiên: vì sao không sinh thẳng từng mẫu?
- Thang mel nén 513 vạch xuống 80 dải mà gần như không mất gì tai nghe được.
- Ba loại vật liệu có ba cấu trúc khác hẳn nhau trên phổ mel.

Bốn cách biểu diễn cùng một giây tiếng

- Dạng sóng thô, phổ độ lớn, phổ mel và token.
- Câu trả lời cho việc không sinh thẳng sóng nằm ở số lượng, không ở chất lượng.
- Bitrate của token thấp hơn PCM (Pulse-Code Modulation – điều chế mã xung) hàng chục lần.



Hình 13.3 — Bốn cách biểu diễn cùng một giây tiếng, và bitrate của chúng

Số token và bitrate

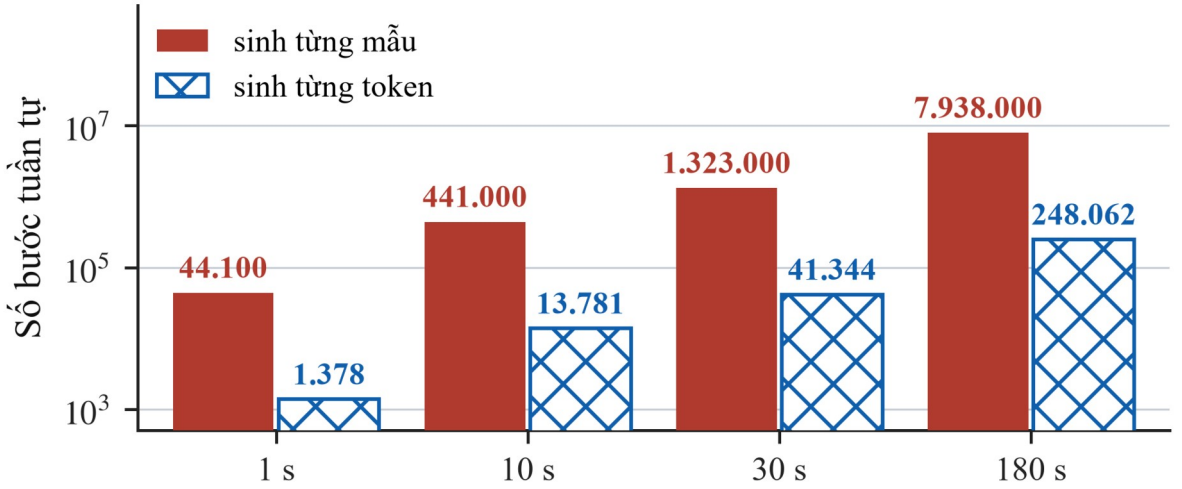
$$R_{\text{token}} = T \cdot \frac{f_s}{H}, \qquad B = T \log_2 K \cdot \frac{f_s}{H}$$

(13.1)

- T là số tầng lượng tử hoá, K là số từ mã mỗi tầng.
- f là tần số lấy mẫu, H là bước nhảy của phép biến đổi Fourier thời gian ngắn.
- T = 8, K = 1.024, H = 256 mẫu ở 44,1 kHz: 1.378 token mỗi giây.
- Bitrate tương ứng 13,78 kbit/s, so với 705,6 kbit/s của PCM mười sáu bit.

Vì sao token là bắt buộc

- Ba phút nhạc là 7.938.000 mẫu.
- Sinh từng mẫu là bảy triệu chín trăm nghìn bước tuần tự, không song song hoá được.
- Cùng ba phút ấy chỉ là 248.062 token, ít hơn 32 lần.



(a) Số bước tuần tự theo độ dài đoạn nhạc

BA PHÚT, SINH TỪNG MẪU

7.938.000

bước tuần tự, không song song hoá được

BA PHÚT, SINH TỪNG TOKEN

248.062

ít hơn 32 lần

(b) Ba phút nhạc: hai cách sinh

Hình 13.4 — Cái giá tuần tự của tự hồi quy, và vì sao token là bắt buộc

Công thức đổi hertz sang mel

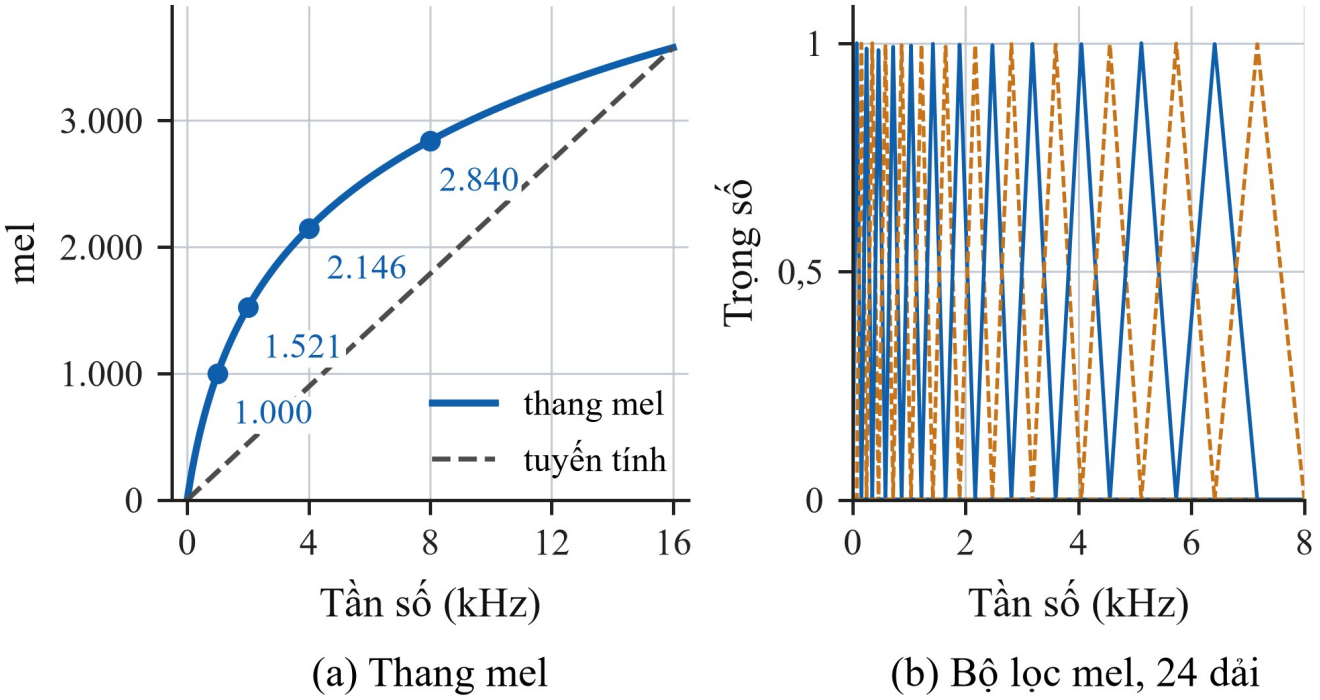
$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

(13.2)

- Công thức không phải quy ước tuỳ ý: nó dựng lại cách tai phân giải tần số (mục 1.7).
- Vùng thấp: hai tần số cách nhau vài chục hertz đã phân biệt được.
- Vùng cao: phải cách nhau hàng trăm hertz mới nghe ra.
- Thang mel đúng là hình dạng ấy.

Thang mel và bộ lọc mel dựng từ nó

- Gộp 513 vạch tần số của một khung thành 80 dải.
- Mỗi dải là một tam giác trên trục tần số, rộng dần về phía cao.
- Gộp thành 80 vạch cách đều thì làm sai lệch cả vùng thấp.



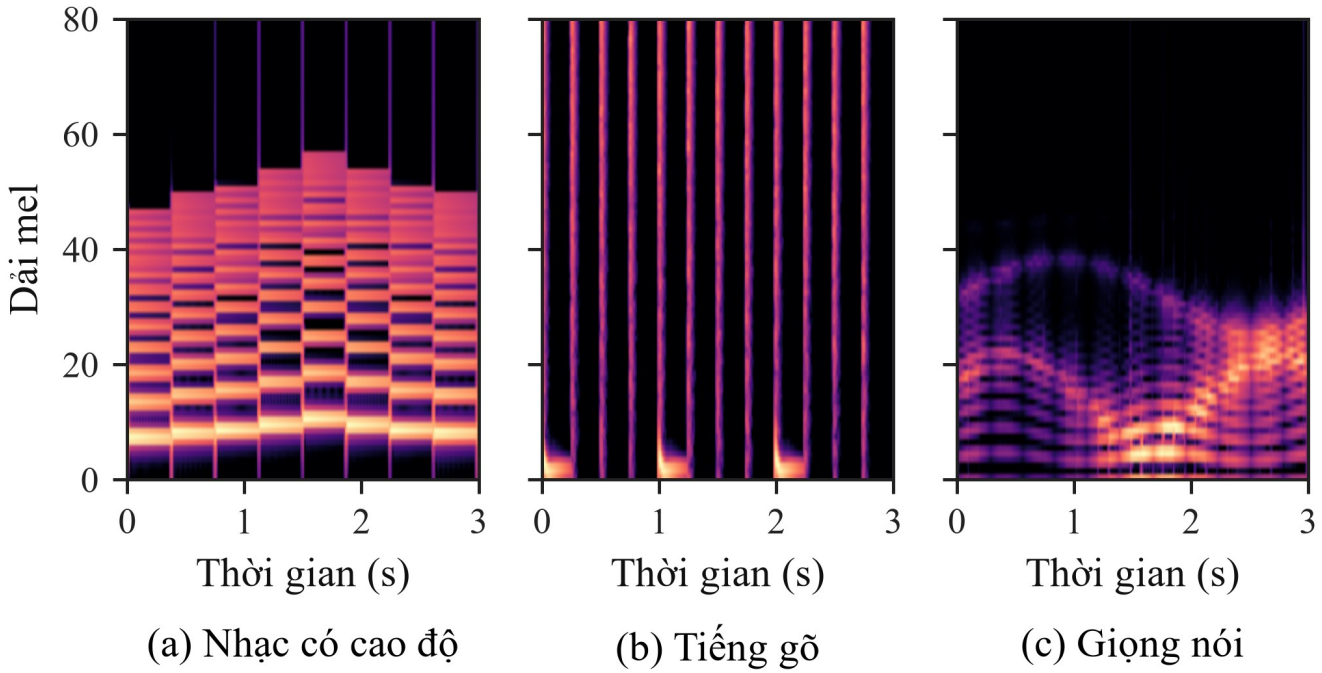
Hình 13.5 — Thang mel, và bộ lọc mel dựng từ nó

Nén mel không miễn phí

Phép chiếu xuống chiều thấp không đảo ngược chính xác (mục 13.6.3). Phổ mel chỉ giữ độ lớn, bỏ hẳn pha: mô hình sinh xong phổ mel vẫn chưa có tiếng, nên bắt buộc phải có tầng ba.

Phổ mel của nhạc, tiếng gõ và giọng nói

- Cấu trúc mắt nhìn ra ở đây là cấu trúc mọi mô hình phải học.
- Mục 13.7 khai thác chính cấu trúc này để tách nguồn mà không cần học.
- Giỏi nhạc có cao độ chưa chắc giỏi tiếng gõ.



Hình 13.6 — Phổ mel của nhạc có cao độ, tiếng gõ, và giọng nói

Bốn cách biểu diễn để tra cứu

Cách biểu diễn	Số mỗi giây	Giữ pha	Dùng làm đích cho
Dạng sóng thô	44.100	Có	Không mô hình thật nào
Phổ độ lớn 513 vạch	88.372	Không	Ít dùng, quá nhiều số
Phổ mel 80 dải	13.781	Không	Họ khuếch tán
Token 8 tầng × 1.024 mã	1.378	Gián tiếp	Họ tự hồi quy

MỤC 13.3

Codec nơ-ron — nơi token của nhạc sinh ra

- Lượng tử hoá véc-tơ một tầng bị gãy vì số từ mã tăng theo hàm mũ.
- Lượng tử hoá dư chồng nhiều mã sách nhỏ lên nhau.
- Chương đo chất lượng theo cách chia tầng và theo số tầng, kể cả việc bỏ tầng khi chạy.

Số từ mã theo số bit

$$N_{\text{từ mã}} = T \cdot 2^{B/T}$$

(13.3)

- Số từ mã K = 2 mũ b, với b là số bit mỗi khung.
- Đây là gốc của mọi khó khăn của cách một tầng.

Bộ nhớ, tốc độ và dữ liệu

Bộ nhớ: mã sách 65.536 từ, 40 chiều là hơn 10 megabyte. Tốc độ: ở 172 khung mỗi giây là hơn mười một triệu phép so mỗi giây. Dữ liệu: học không đủ khung thì mã sách chỉ thuộc lòng (mục 13.3.3).

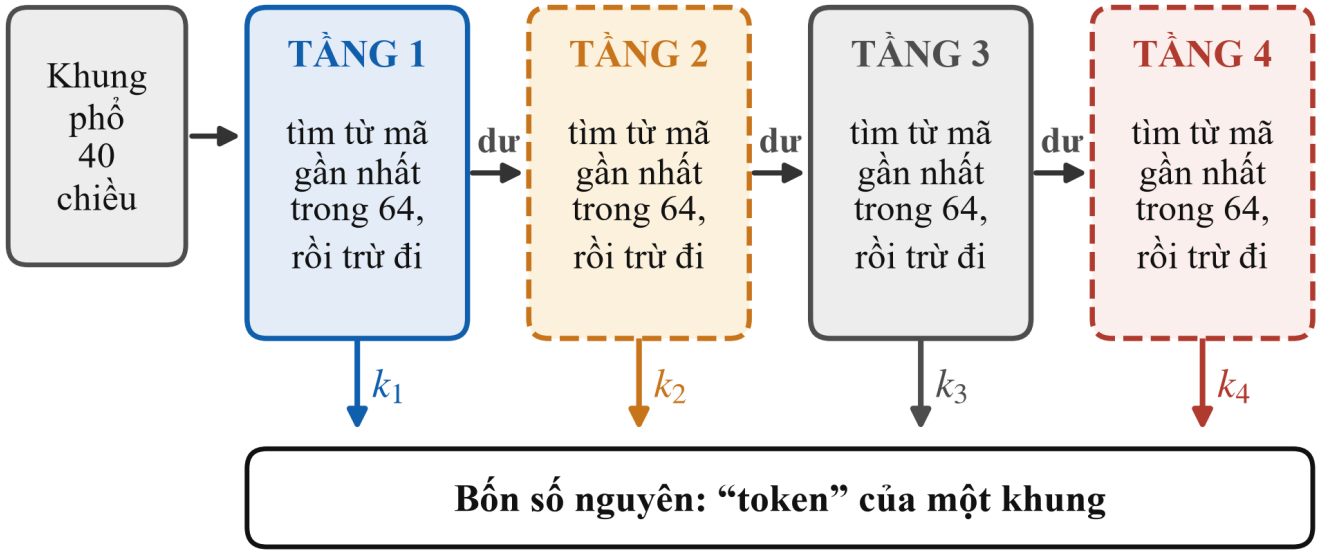
Phần dư truyền từ tầng này sang tầng sau

$$\mathbf{x} \approx \sum_{t=1}^{T} \mathbf{c}_t[k_t], \qquad \mathbf{r}_t = \mathbf{r}_{t-1} - \mathbf{c}_t[k_t]$$

(13.4)

- Mỗi tầng chỉ phải dựng phần mà các tầng trước còn thiếu.
- Tổng các từ mã được chọn ở mọi tầng là véc-tơ tái dựng.

Mỗi tầng dựng phần dư của tầng trước



Tầng một dựng phần lớn; mỗi tầng sau chỉ dựng phần dư mà các tầng trước còn để lại.

Bỏ bớt tầng cuối vẫn nghe được, chỉ kém đi (công thức 13.3); codec đổi bitrate mà không phải học lại. Mục 13.3.4 đo điều đó.

Hình 13.7 — Sơ đồ lượng tử hoá dư: mỗi tầng dựng phần dư của tầng trước

Học chồng mã sách

```
def rvq_hoc(X, so_ma, so_tang, vong):
    du = X.copy()
    sachs = []
    for t in range(so_tang):
        C = hoc_ma_sach(du, so_ma, vong)   # k-means
        sachs.append(C)
        _, dung = luong_tu_vector(du, C)
        du = du - dung                     # phần DƯ đi tiếp
    return sachs
```

Trong codec thật, phép biến đổi trước và sau khối này là mạng nơ-ron; phần lượng tử hoá thì đúng như trên.

SNR tái dựng

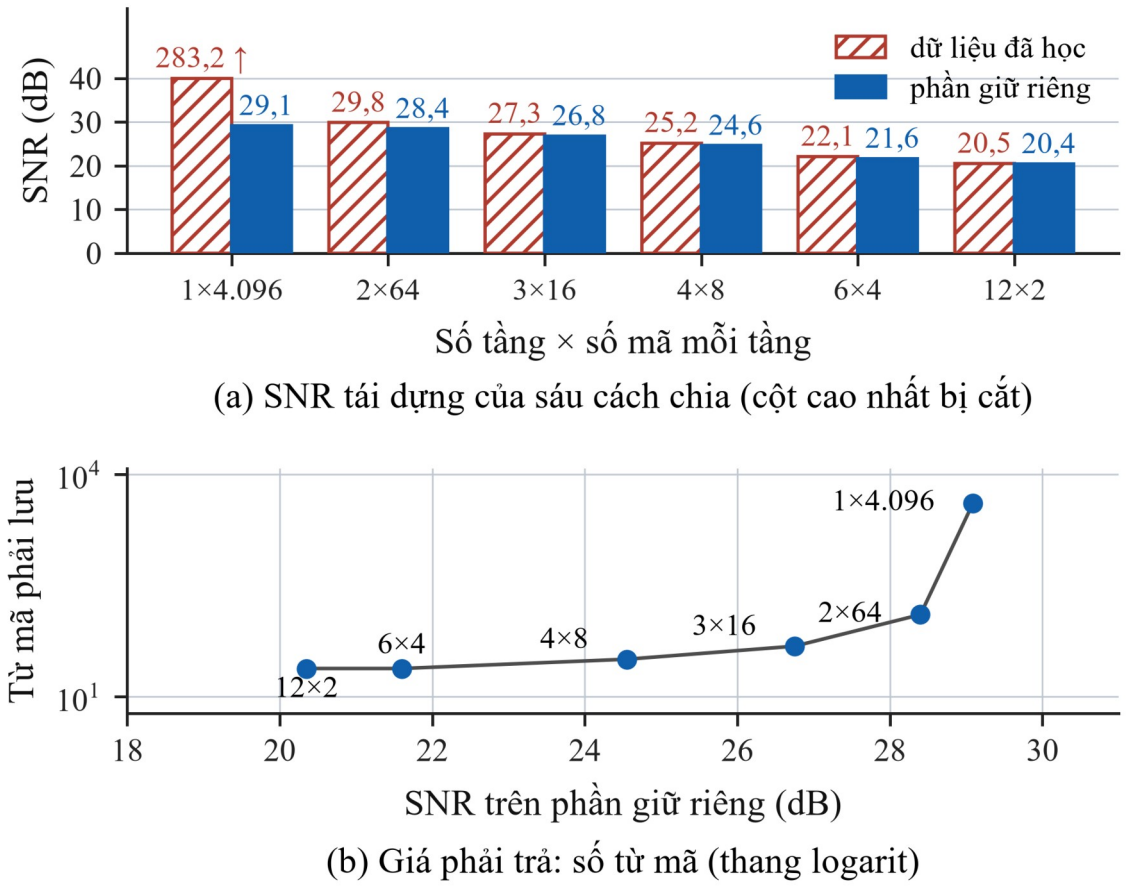
$$\mathrm{SNR}_{\text{tái dựng}} = 10 \log_{10} \frac{\sum_n \|\mathbf{x}_n\|^2}{\sum_n \|\mathbf{x}_n - \hat{\mathbf{x}}_n\|^2}$$

(13.5)

- Tỉ số giữa năng lượng tín hiệu và năng lượng sai số tái dựng, tính bằng decibel.
- Cùng định nghĩa SNR (Signal-to-Noise Ratio – tỉ số tín hiệu trên nhiễu) đã dùng từ mục 3.5.

Sáu cách chia mười hai bit mỗi khung

- Sáu cách đều tiêu đúng mười hai bit nên cùng bitrate.
- Câu hỏi thiết kế: nên chia thành mấy tầng?
- Phải đọc hai cột SNR cùng nhau.



Hình 13.8 — Sáu cách chia cùng một ngân sách bit, và cái giá của mỗi cách

Sáu cách chia mười hai bit mỗi khung

Cách chia	SNR đã học	SNR giữ riêng	Từ mã phải lưu
1 tầng × 4.096 mã	283,2 dB	29,09 dB	4.096
2 tầng × 64 mã	29,84 dB	28,40 dB	128
3 tầng × 16 mã	27,29 dB	26,75 dB	48
4 tầng × 8 mã	25,18 dB	24,55 dB	32
6 tầng × 4 mã	22,07 dB	21,60 dB	24
12 tầng × 2 mã	20,46 dB	20,35 dB	24

283 decibel là con số vô nghĩa

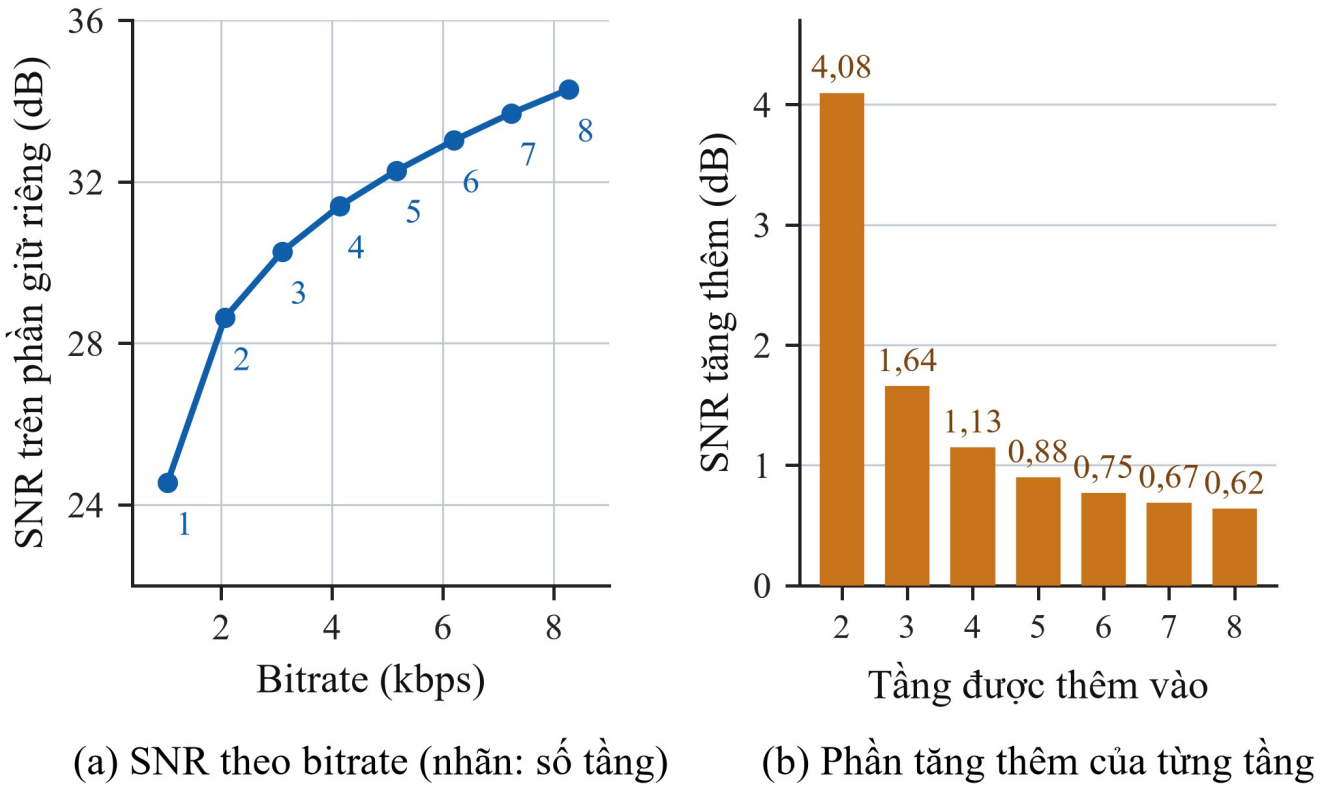
Mã sách có 4.096 từ trong khi chỉ có 2.586 khung học, nên nó thuộc lòng từng khung. Trên phần giữ riêng chỉ còn 29,09 dB. Phép đo chỉ có nghĩa khi làm trên dữ liệu chưa học.

Hai tầng gần bằng một tầng, với 1/32 số từ mã

Trên phần giữ riêng, một tầng chỉ hơn hai tầng 0,69 dB, nhưng phải lưu và duyệt nhiều gấp 32 lần từ mã. Chia quá nhỏ cũng kém: mười hai tầng hai mã thấp hơn hai tầng tới tám decibel. Chọn tốt nhất: hai đến bốn tầng.

Thêm tầng thì được thêm bao nhiêu decibel

- Giữ nguyên số mã mỗi tầng rồi tăng số tầng.
- Tầng hai thêm 4,08 dB; tầng tám chỉ thêm 0,62 dB.
- Bitrate tăng gấp tám mà chất lượng chỉ tăng 9,7 dB.

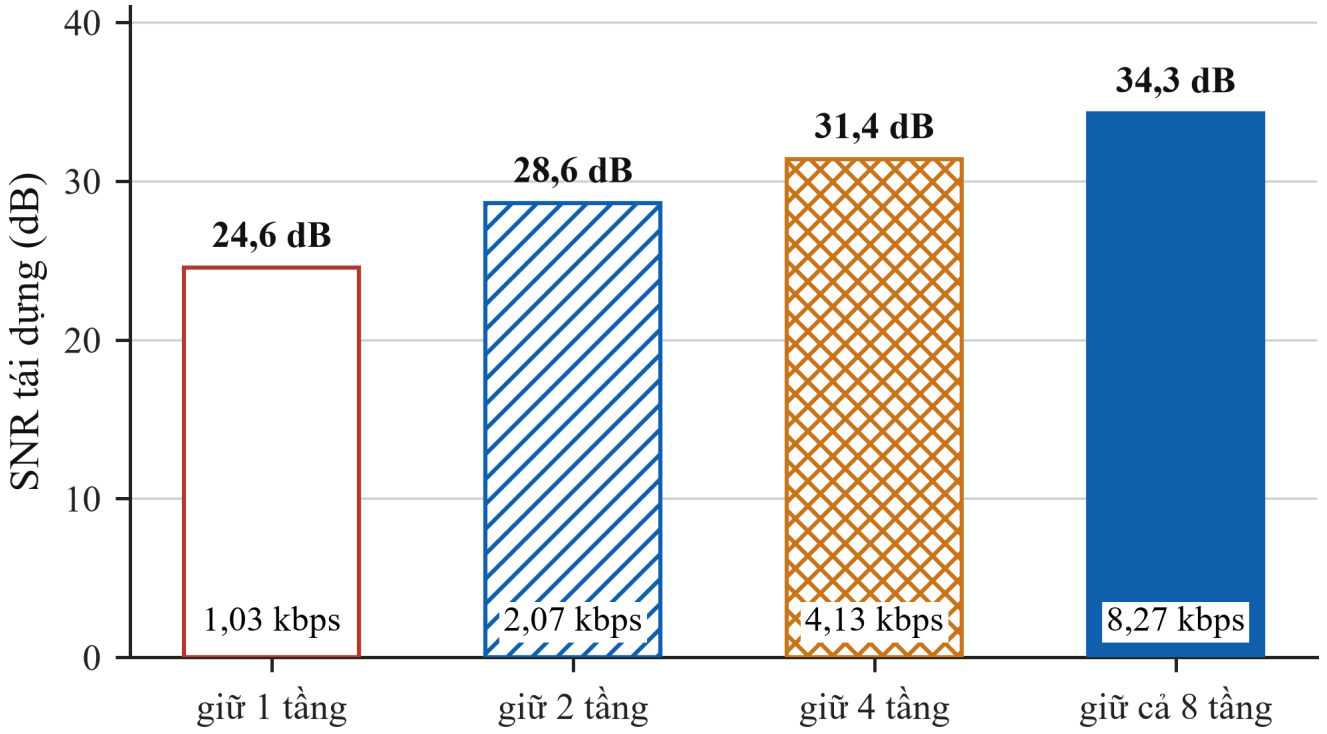


(a) SNR theo bitrate (nhãn: số tầng) (b) Phần tăng thêm của từng tầng

Hình 13.9 — Thêm tầng thì được thêm bao nhiêu decibel

Đọc bao nhiêu tầng tuỳ người dùng

- Bốn mức chất lượng dựng từ đúng một chuỗi token.
- Không mã hoá lại, không học lại, không lưu bốn bản.
- Bộ giải mã chỉ cần đọc ít tầng hơn.



Hình 13.10 — Cùng một chuỗi token, đọc bao nhiêu tầng thì tuỳ người dùng

MỤC 13.4

Họ mô hình tự hồi quy

- Sinh chuỗi từng phần tử, mỗi phần tử phụ thuộc phần tử đã sinh.
- Entropy chéo đo trực tiếp mô hình đã học được bao nhiêu.
- Ba tham số lấy mẫu quyết định chuỗi đa dạng hay suy sụp thành vòng lặp.

Entropy chéo của mô hình tự hồi quy

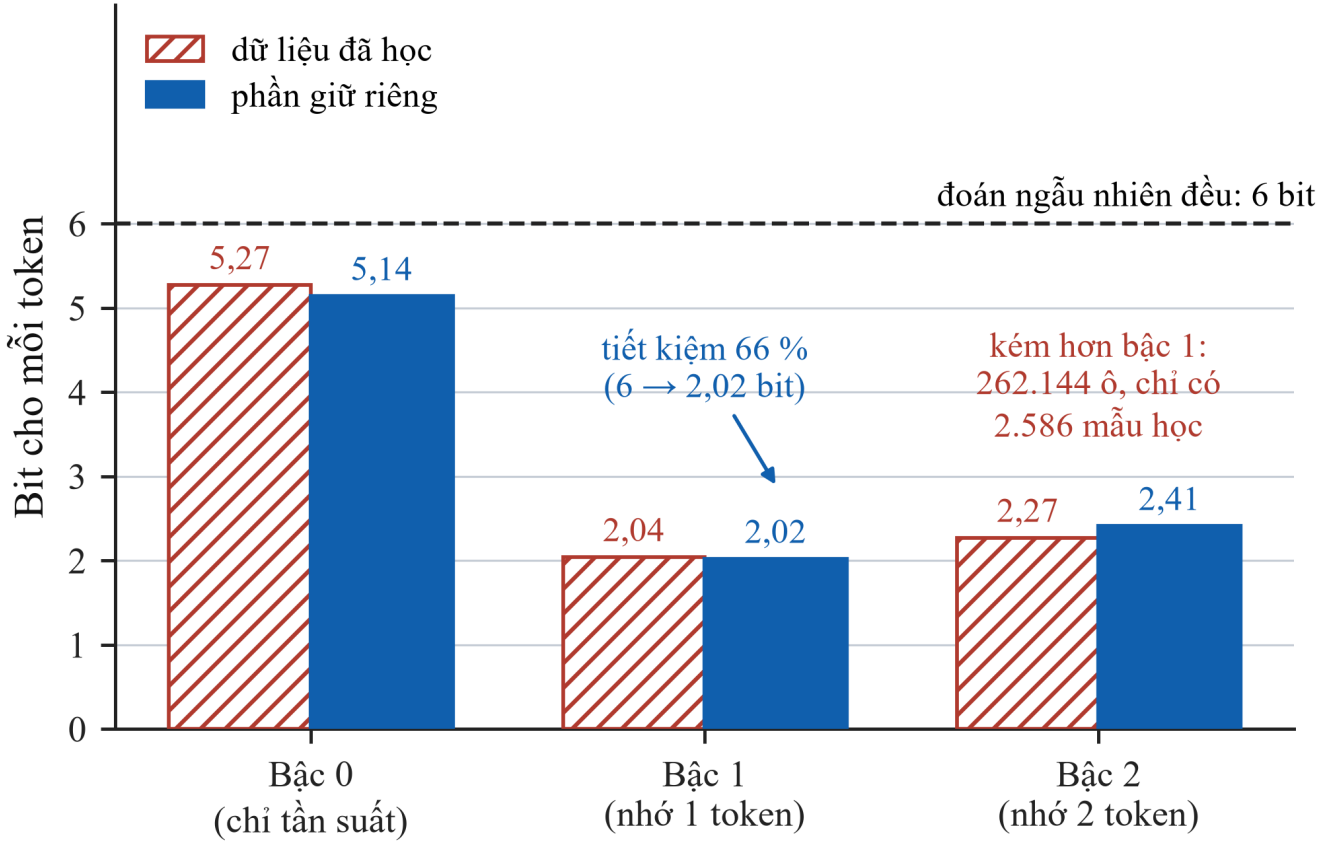
$$p(k_{1..N}) = \prod_{i=1}^{N} p(k_i \mid k_{<i}), \qquad \bar{H} = -\frac{1}{N} \sum_{i=1}^{N} \log_2 p(k_i \mid k_{<i})$$

(13.6)

- Xác suất của chuỗi là tích các xác suất có điều kiện theo phần tử trước.
- Vế thứ hai là entropy chéo: số bit trung bình để ghi một token.
- Vừa là hàm mất mát khi huấn luyện, vừa là thước đo mức đã học.

Ngữ cảnh đáng bao nhiêu bit

- Đo ngay trên chuỗi token thật của mục 13.3.
- Nhớ một token tiết kiệm 66 phần trăm so với đoán ngẫu nhiên.
- Nhớ hai token lại kém hơn vì thiếu dữ liệu.



Hình 13.11 — Ngữ cảnh đáng bao nhiêu bit, và vì sao nó không miễn phí

Entropy chéo theo độ dài ngữ cảnh

Mô hình	Đã học	Giữ riêng	Nhận xét
Đoán ngẫu nhiên đều	6,00 bit	6,00 bit	Trần trên, log2 của 64
Bậc 0	5,27 bit	5,14 bit	Chỉ biết tần suất
Bậc 1	2,04 bit	2,02 bit	Tiết kiệm 66 %
Bậc 2	2,27 bit	2,41 bit	Kém hơn bậc 1: thiếu dữ liệu

Ngữ cảnh dài hơn đòi hỏi dữ liệu nhiều hơn

Bậc một có 4.096 ô với 2.586 mẫu học; bậc hai có 262.144 ô với cùng số mẫu, trung bình chưa tới một phần trăm ô có mẫu. Mô hình lớn chỉ đẩy điểm gãy đi xa hơn bằng cách chia sẻ tham số giữa các ngữ cảnh giống nhau.

Nhiệt độ áp lên phân phối

$$p_\tau(k) = \frac{p(k)^{1/\tau}}{\sum_j p(j)^{1/\tau}}$$

(13.7)

- Chọn token xác suất cao nhất cho chuỗi lặp nhàm chán.
- Lấy mẫu đúng phân phối thì đa dạng nhưng dễ lạc.
- Nhiệt độ nhỏ hơn 1 làm phân phối nhọn: an toàn nhưng lặp.
- Top-k giữ k xác suất lớn nhất; top-p giữ nhóm nhỏ nhất có tổng vượt ngưỡng.

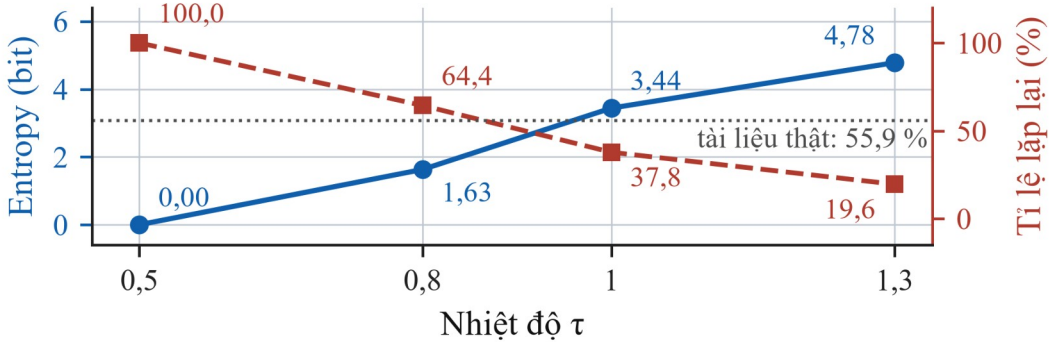
Ba phép lọc phân phối

```
def ap_nhiet(p, nhiet):
    q = log(maximum(p, 1e-30)) / nhiet
    q = exp(q - q.max())
    return q / q.sum()

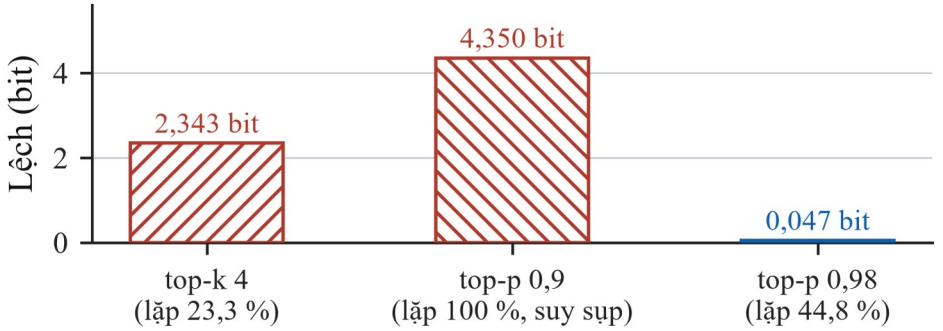
def loc_top_k(p, k):
    nguong = sort(p)[-k]
    q = where(p >= nguong, p, 0.0)
    return q / q.sum()

def loc_top_p(p, nguong):
    thu = argsort(p)[::-1]
    giu = searchsorted(cumsum(p[thu]), nguong) + 1
    q = zeros_like(p); q[thu[:giu]] = p[thu[:giu]]
    return q / q.sum()
```

Nhiệt độ, top-k và top-p trên chuỗi token của bài



(a) Nhiệt độ: entropy (liền) và tỉ lệ lặp lại (đứt)



(b) Top-k và top-p: độ lệch so với tài liệu thật

Hình 13.12 — Nhiệt độ, top-k và top-p đo trên chính chuỗi token của bài

Tám cách lấy mẫu, đo trên 3.000 token

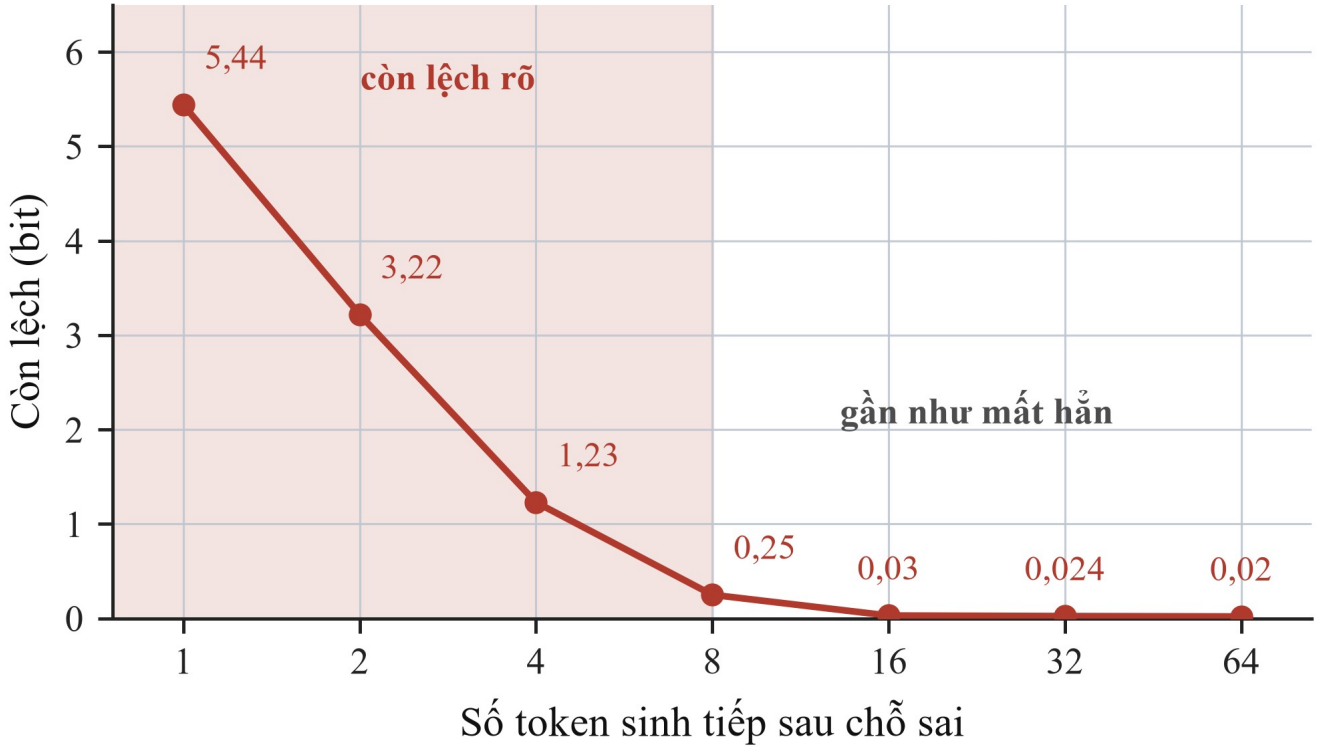
Cách lấy mẫu	Entropy	Tỉ lệ lặp	Lệch tài liệu thật
Tài liệu thật	5,20 bit	55,9 %	—
Nhiệt 0,5	0,00 bit	100,0 %	4,350 bit
Nhiệt 0,8	1,63 bit	64,4 %	0,179 bit
Nhiệt 1,0	3,44 bit	37,8 %	0,108 bit
Nhiệt 1,3	4,78 bit	19,6 %	0,354 bit
Top-k 4	1,85 bit	23,3 %	2,343 bit
Top-p 0,9	0,00 bit	100,0 %	4,350 bit
Top-p 0,98	2,93 bit	44,8 %	0,047 bit

Nhiệt 0,5 và top-p 0,9 lặp đúng một token

Tỉ lệ lặp 100 phần trăm: nguyên nhân là tham số, không phải mô hình. Top-p không tự động an toàn hơn top-k; ngưỡng vẫn phải chọn bằng cách đo. Lệch thấp nhất là top-p 0,98, thấp hơn cả lấy mẫu nguyên bản.

Một token sai kéo dài bao lâu

- Huấn luyện luôn thấy ngữ cảnh đúng; khi sinh phải thấy ngữ cảnh do chính mô hình sinh.
- Mô hình bậc một: sau tám token còn lệch 0,25 bit.
- Sau ba mươi hai token chỉ còn 0,024 bit.



Hình 13.13 — Một token sai kéo dài bao lâu trong mô hình bậc một

Chương dừng ở mức mô tả chỗ này

Sai lệch mất nhanh chỉ vì mô hình bậc một không có trí nhớ dài. Mô hình có cửa sổ ngữ cảnh dài giữ token sai trong ngữ cảnh: đó mới là lệch phơi nhiễm thật, và mô hình đồ chơi không dựng lại được.

MỤC 13.5

Họ mô hình khuếch tán

- Bắt đầu từ nhiễu thuần và gọt dần thành tín hiệu; sinh cả đoạn mỗi bước.
- Quá trình thuận biết trước hoàn toàn, nên dữ liệu huấn luyện sinh miễn phí.
- Lịch nhiễu và số bước lấy mẫu quyết định chất lượng và tốc độ.

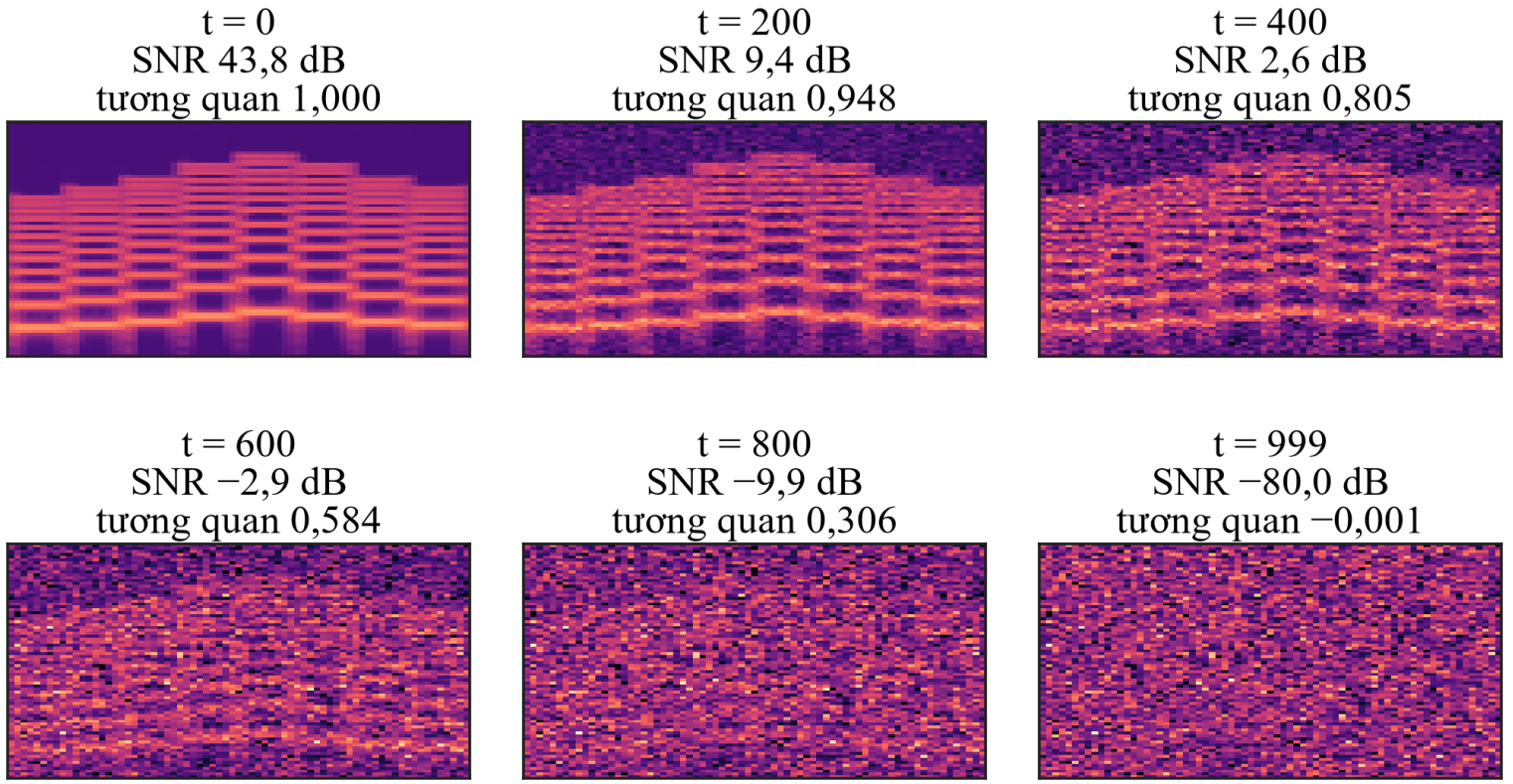
Trộn nhiễu theo tỉ lệ đã biết

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathrm{N}(0, \mathbf{I})$$

(13.8)

- alpha ngang là phần tín hiệu còn lại ở bước t; epsilon là nhiễu Gauss chuẩn.
- Chỉ là một phép trộn có tỉ lệ đã biết.
- Huấn luyện: chọn bước t bất kỳ, trộn nhiễu, bảo mô hình đoán lại nhiễu đã thêm.

Quá trình thuận chạy trên phổ mel của một giai điệu



Hình 13.15 — Quá trình thuận chạy trên phổ mel thật của một giai điệu

Alpha ngang theo lịch tuyến tính

$$\mathrm{SNR}(t) = 10 \log_{10} \frac{\bar{\alpha}_t}{1 - \bar{\alpha}_t}$$

(13.9)

- Alpha ngang đi từ gần một xuống gần không theo lịch do người thiết kế chọn.
- Lịch tuyến tính là lựa chọn đầu tiên trong lịch sử.
- Vấn đề chỉ lộ ra khi nhìn bằng decibel.

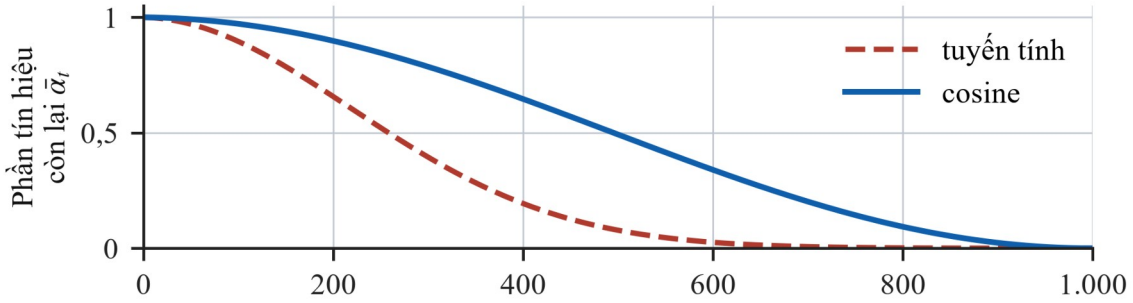
Lịch cosine

$$\bar{\alpha}_t = \frac{f(t)}{f(0)}, \qquad f(t) = \cos^2\left(\frac{t/T + s}{1+s} \cdot \frac{\pi}{2}\right)$$

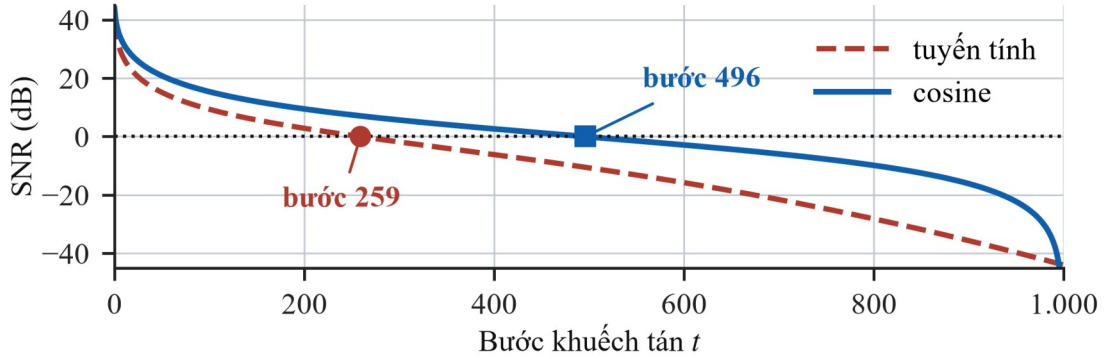
(13.10)

- Lịch tuyến tính xuống dưới 0 dB ngay ở bước 259 trên một nghìn.
- Hơn 74 phần trăm số bước còn lại khử thứ mà tín hiệu đã yếu hơn nhiễu.
- Lịch cosine đẩy mốc ấy tới bước 496, đúng giữa hành trình.
- Không thêm tham số học nào; chỉ đổi công thức.

Hai lịch nhiễu và SNR của chúng



(a) Hai lịch nhiễu: phần tín hiệu còn lại theo bước



(b) Tỉ số tín hiệu trên nhiễu của hai lịch

Hình 13.16 — Hai lịch nhiễu, và tỉ số tín hiệu trên nhiễu của chúng

Đo bằng bài toán đồ chơi để tách riêng số bước

Mô hình thật không tách được lỗi do ít bước với lỗi do học kém. Với hỗn hợp ba Gauss một chiều, bộ khử nhiễu tối ưu viết ra bằng công thức. Chuyển được sang bài toán thật là hình dạng đường cong, không phải con số tuyệt đối.

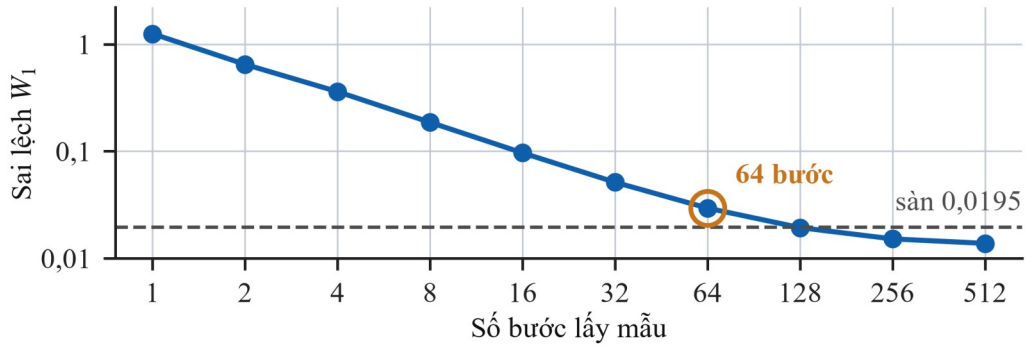
Một bước lấy mẫu kiểu ddim

$$\mathbf{x}_s = \sqrt{\bar{\alpha}_s}\,\hat{\mathbf{x}}_0 + \sqrt{1-\bar{\alpha}_s}\,\hat{\boldsymbol{\varepsilon}}_t$$

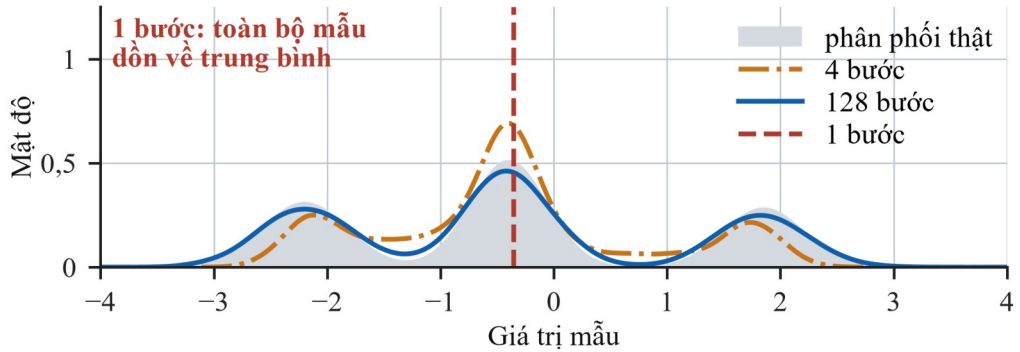
(13.11)

- Từ trạng thái ở bước t, ước lượng bản sạch.
- Rồi nhảy thẳng tới bước s bất kỳ nhỏ hơn t.
- Số bước dùng thật tách rời khỏi số bước đã huấn luyện.

Số bước lấy mẫu đổi chất lượng bao nhiêu



(a) Số bước đổi chất lượng bao nhiêu



(b) Ít bước thì mất gì: phân phối sinh ra

Hình 13.17 — Số bước lấy mẫu đổi chất lượng bao nhiêu, và ít bước thì mất gì

Sai lệch phân phối theo số bước lấy mẫu

Số bước	Sai lệch W1	So với sàn	Nhận xét
1	1,270	65 lần	Mọi mẫu dồn về trung bình
4	0,344	18 lần	Có ba đỉnh, sai tỉ lệ
16	0,090	4,6 lần	Đã gần đúng
64	0,024	1,2 lần	Đủ dùng
128	0,020	1,0 lần	Chạm sàn
512	0,020	1,0 lần	Không hơn 128 bước

Gấp đôi số bước, sai lệch giảm gần một nửa

- Sai lệch giảm đều cho tới khi chạm sàn ở khoảng 64 đến 128 bước.
- Hệ thống thật dùng vài chục bước, không dùng cả nghìn.
- Hệ thống một bước phải được chưng cất riêng.
- Một bước cho trung bình của phân phối: tiếng nghe mờ và thiếu chi tiết, không méo hay nhiễu.

Tự hồi quy so với khuếch tán

- Ba mươi giây nhạc: 41.344 bước tuần tự với mô hình tự hồi quy tám tầng.
- Khoảng năm mươi bước với mô hình khuếch tán, bất kể độ dài.
- Tỉ số 827 lần, tăng tuyến tính theo độ dài.

Tiêu chí	Tự hồi quy trên token	Khuếch tán trên phổ mel
Số bước tuần tự cho 30 giây	41.344 bước, tăng theo độ dài (gấp 827 lần)	50 bước, không phụ thuộc độ dài
Sinh được đoạn dài không	Được, cứ nối tiếp mãi	Khó: cửa sổ cố định, phải ghép
Sửa giữa chừng được không	Không: đã sinh là xong	Được: sửa mặt nạ rồi khử nhiễu lại
Đích sinh ra là gì	Chuỗi token rời rạc	Phổ mel liên tục
Cần gì ở tầng ba	Bộ giải mã của codec	Một vocoder, chỗ mất mát (mục 13.6)
Cái sai điển hình	Suy sụp thành vòng lặp, hoặc lạc điệu	Mờ, mất chi tiết khi ít bước

Hai họ đang hội tụ: nhiều hệ thống mới sinh token bằng khuếch tán, hoặc sinh phổ mel bằng mô hình tự hồi quy trên các mảnh phổ.

Bảng này so sánh hai cơ chế, không phải hai nhóm sản phẩm.

Hình 13.18 — Tự hồi quy so với khuếch tán: sáu tiêu chí

Hai câu hỏi riêng biệt

Bảng so sánh là về hai cơ chế, không phải hai sản phẩm. Nhiều hệ thống sinh token bằng khuếch tán; nhiều hệ thống khác sinh phổ mel bằng tự hồi quy. Hỏi: nó sinh ra cái gì, và nó sinh theo cơ chế nào.

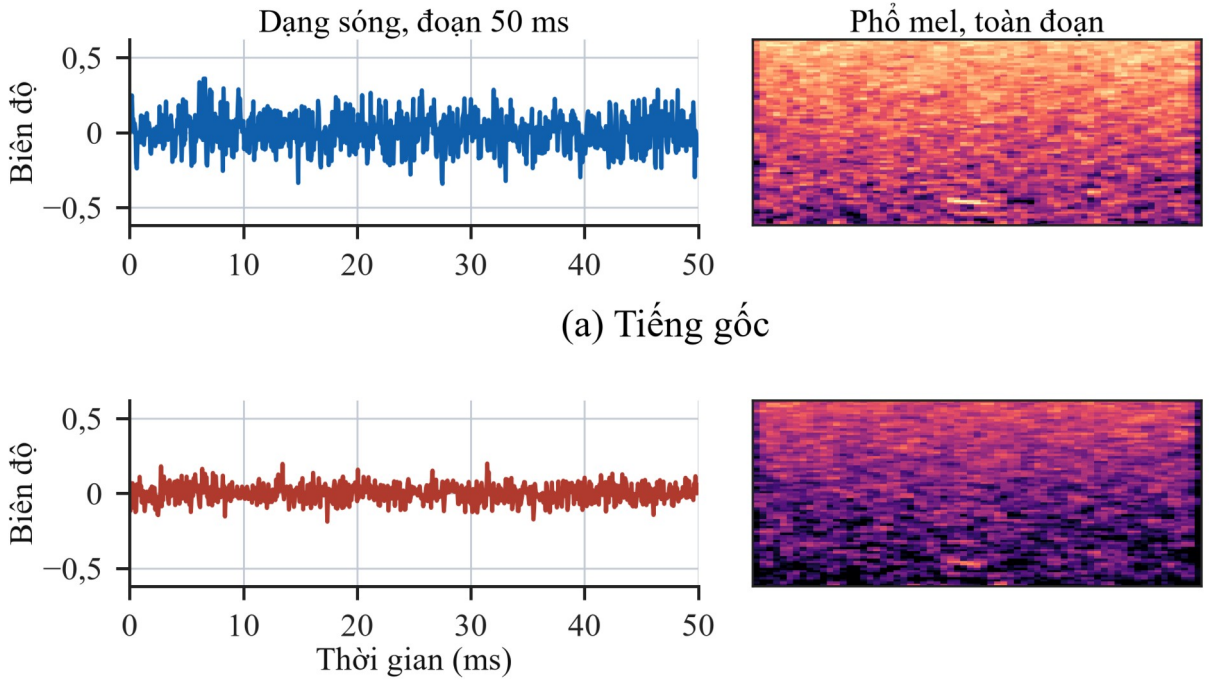
MỤC 13.6

Giọng, vocoder và bài toán pha

- Phổ mel hoàn hảo vẫn chưa phải là tiếng.
- Griffin-Lim giảm sai lệch độ lớn nhưng không đưa dạng sóng về gần bản gốc.
- Cổ chai mel gây mất mát trước cả khi nói tới pha.

Cùng độ lớn phổ, hai pha, hai tiếng khác hẳn

- Phổ độ lớn loại bỏ pha: giữ độ lớn, bỏ thông tin về thời điểm.
- Pha quyết định dạng sóng, không phải chi tiết nhỏ.
- Bản thứ hai nghe là tiếng rè, không còn nhận ra giọng.



Hình 13.19 — Cùng một độ lớn phổ, hai pha khác nhau, hai tiếng khác hẳn

Hai bản cùng độ lớn phổ, đo bằng ba thước

Thước đo	Giá trị	Nói lên điều gì
Khoảng cách mel	6,98 dB	Gần như không thấy khác biệt
Khoảng cách phổ log	11,30 dB	Cũng không thấy gì đáng kể
SI-SDR	−39,4 dB	Hoàn toàn không phải cùng một tiếng

Mọi thước đo dựa trên độ lớn phổ đều bảo hai bản gần như giống nhau.

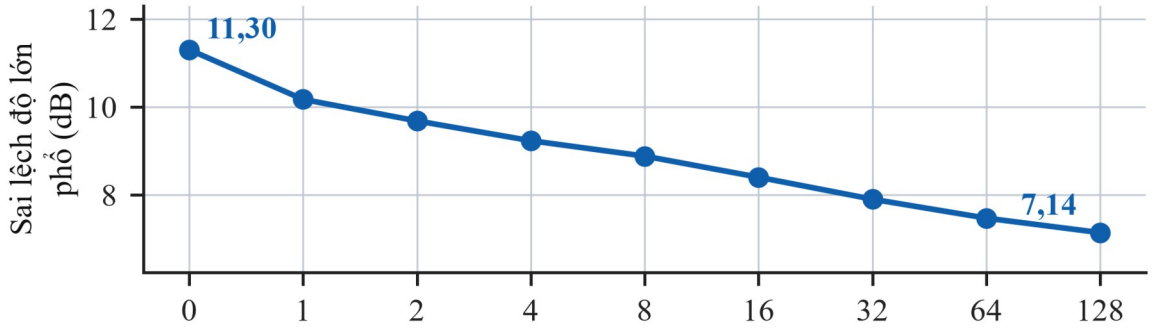
Griffin-Lim

```
def griffin_lim(bien_do, vong):
    pha = exp(1j * uniform(-pi, pi, bien_do.shape))
    X = bien_do * pha              # xuất phát từ pha ngẫu nhiên
    for _ in range(vong):
        y = istft(X)               # về miền thời gian
        Y = stft(y)                # rồi quay lại miền phổ
        X = bien_do * exp(1j * angle(Y))   # GIỮ pha, THAY độ lớn
    return istft(X)
```

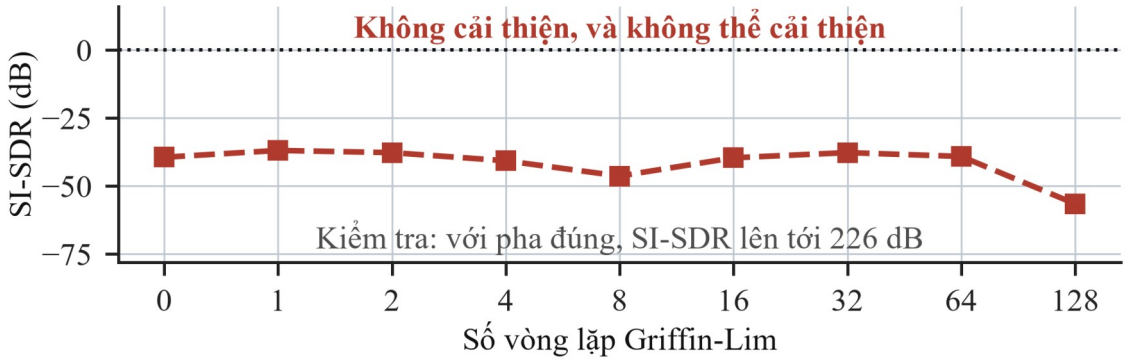
Vòng lặp giữ độ lớn, đổi pha.

Griffin-Lim cải thiện gì và không cải thiện gì

- Sai lệch độ lớn phổ giảm từ 11,30 xuống 7,14 dB sau 128 vòng, đều và đơn điệu.
- Dạng sóng dựng ra không tiến gần tiếng gốc.
- Hai hình phải đọc cùng nhau.



(a) Cái mà Griffin-Lim cải thiện: độ lớn phổ



(b) Cái mà nó không cải thiện: dạng sóng

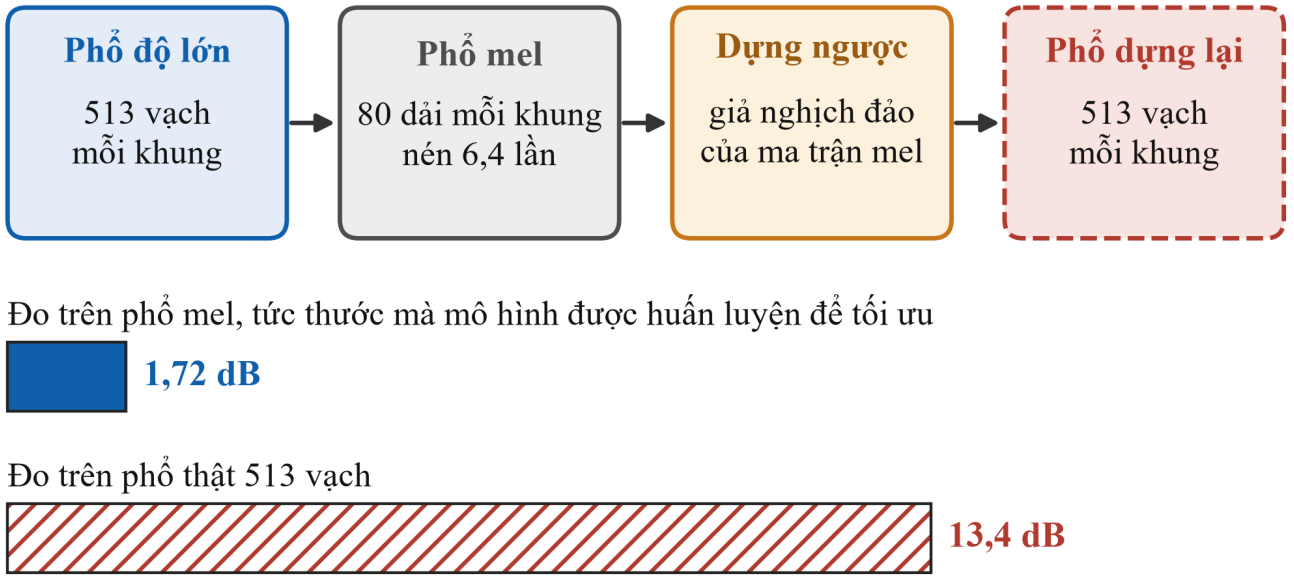
Hình 13.20 — Cái mà Griffin-Lim cải thiện, và cái mà nó không cải thiện

Có vô số dạng sóng cùng một độ lớn phổ

Griffin-Lim hội tụ về một dạng sóng trong số đó, không nhất thiết là bản gốc. Không dùng SI-SDR (Scale-Invariant Signal-to-Distortion Ratio – tỉ số tín hiệu trên méo bất biến tỉ lệ) để chấm bộ dựng pha. Có pha đúng thì SI-SDR lên tới 226 dB, nên vấn đề không ở công cụ đo. Đầu ra nghe có chất kim loại đặc trưng.

Đo trên mel thì ổn, đo trên phổ thật thì không

- Nén 513 vạch xuống 80 dải là phép chiếu nên không đảo ngược chính xác.
- Cùng một bản dựng lại: lệch 1,72 dB theo khoảng cách mel.
- Cũng bản ấy, lệch 13,4 dB theo phổ thật.



Cùng một bản dựng lại, nhưng sai lệch lớn gấp 7,8 lần khi đo trên phổ thật.

Vì vậy vocoder hiện đại học cách đi từ mel về sóng, không dựng ngược ma trận mel; và mô hình sinh phổ mel không nên chấm bằng chính khoảng cách mel của nó.

Hình 13.21 — Cổ chai mel: đo trên mel thì ổn, đo trên phổ thật thì không

Vocoder học và quy tắc chấm điểm

Vocoder hiện đại học ánh xạ từ mel về sóng nên bù được thông tin mà phép chiếu đã bỏ. Mô hình sinh phổ mel không nên được đánh giá bằng chính khoảng cách mel của nó: đó là chấm bài bằng chính đề bài.

Ba khối của hệ thống sinh giọng

Khối	Việc	Sai thì nghe ra sao
Từ chữ sang biểu diễn	Văn bản thành chuỗi âm vị, trường độ, ngữ điệu	Đọc sai, ngắt sai, ngữ điệu phẳng
Mô hình âm học	Sinh phổ mel hoặc token	Sai chất giọng; lặp một âm tiết
Vocoder	Từ phổ mel hoặc token về dạng sóng	Chất kim loại, rè, mất phụ âm

Với tiếng Việt, thanh điệu phải được mô hình hoá tường minh ở khối thứ nhất.

MỤC 13.7

Tách nguồn — và hai trần trên không vượt được

- Tách bằng mặt nạ là khuôn chung của gần như mọi phương pháp trên phổ.
- Hai trần trên đều tính được trước khi huấn luyện.
- Hai chuẩn so sánh không cần học: lọc trung vị và NMF (Non-negative Matrix Factorization – phân tích ma trận không âm).

Ước lượng nguồn bằng mặt nạ

$$M_{\text{tỷlệ}}[f,n] = \frac{|S[f,n]|}{|S[f,n]| + |N[f,n]|}$$

(13.12)

- Tách nguồn là bài toán ngược của trộn.
- Ước lượng một mặt nạ trong khoảng từ không đến một cho mỗi ô phổ.
- Nhân mặt nạ với độ lớn của hỗn hợp.

Khuôn chung của mọi phương pháp tách bằng mặt nạ

- Chỉ khối ước lượng mặt nạ khác nhau giữa các phương pháp.
- Bốn khối còn lại hoàn toàn giống nhau.
- Một trong số đó đặt ra giới hạn không phương pháp nào vượt được.



Đây là khối duy nhất cần học, và là khối duy nhất khác nhau giữa các phương pháp.

VÀ ĐÂY LÀ CHỖ ĐẶT RA MỘT TRẦN TRÊN KHÔNG VƯỢT ĐƯỢC

Mặt nạ chỉ sửa độ lớn; pha lấy nguyên của hỗn hợp, tức pha của hai nguồn trộn lẫn. Mặt nạ hoàn hảo tới đâu, phần sai do pha vẫn còn. Mục 13.7.2 đo con số ấy.

Vì vậy các hệ thống tách nguồn hiện đại chuyển sang làm việc trên miền thời gian, nơi không có pha của hỗn hợp để phải vay mượn.

Hình 13.22 — Khuôn chung của mọi phương pháp tách bằng mặt nạ

Trần của mặt nạ và trần do vay pha

- Trần thứ nhất: biết trước cả hai nguồn, mặt nạ tốt nhất cho kết quả bao nhiêu.
- Mặt nạ nhị phân giao mỗi ô cho nguồn to hơn; mặt nạ tỉ lệ chia theo năng lượng.
- Trần thứ hai: pha vay nguyên của hỗn hợp dù độ lớn đúng tới từng ô.

Thước đo SI-SDR

$$\hat{s} = \frac{\langle \hat{y}, s \rangle}{\|s\|^2} s, \qquad \mathrm{SI-SDR} = 10 \log_{10} \frac{\|\hat{s}\|^2}{\|\hat{y} - \hat{s}\|^2}$$

(13.13)

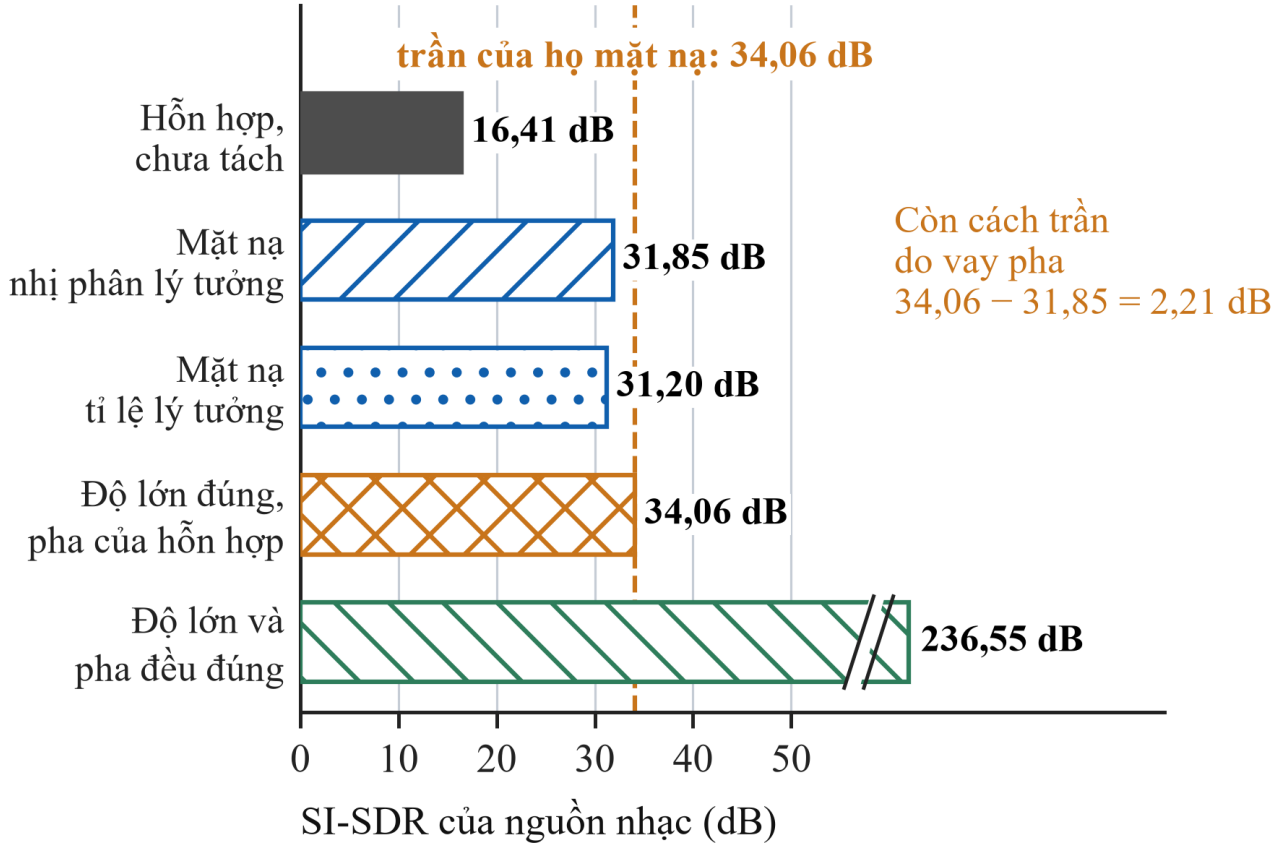
- Bất biến theo tỉ lệ chung nên không bị đánh lừa bởi chênh lệch âm lượng.
- Là thước đo chuẩn của lĩnh vực tách nguồn.

SI-SDR

```
def si_sdr_db(chuan, uoc):
    s = chuan - chuan.mean();  y = uoc - uoc.mean()
    a = dot(y, s) / dot(s, s)        # tỉ lệ tốt nhất — BẤT BIẾN
    chieu = a * s
    nhieu = y - chieu
    return 10 * log10(sum(chieu**2) / sum(nhieu**2))
```

Dòng thứ hai là điều làm nên tên gọi.

Ba trần trên trên cùng một hỗn hợp



Hình 13.23 — Ba trần trên chồng lên nhau, đo trên cùng một hỗn hợp

34,06 dB là trần của họ phương pháp mặt nạ

Không mô hình nào trong họ ấy vượt được, dù lớn đến đâu, dù dữ liệu nhiều đến đâu. Dòng cuối là phép kiểm sức khoẻ: có đủ độ lớn và pha thì dựng lại gần tuyệt đối. Hệ thống hàng đầu vì thế chuyển sang làm trực tiếp trên miền thời gian.

Bốn mức trên cùng một hỗn hợp

Phương pháp	Nguồn nhạc	Nguồn gõ	Cần gì
Hỗn hợp, chưa tách	16,41 dB	−16,25 dB	Không gì
Lọc trung vị	24,17 dB	7,08 dB	Giả thiết cấu trúc; 0,39 giây
NMF, R = 8	30,05 dB	13,43 dB	Không nhãn, phải gán thành phần
Mặt nạ nhị phân lý tưởng	31,85 dB	15,33 dB	Cả hai nguồn

30,05 dB là cận trên, không phải bộ tách mù

Ở đây việc gán thành phần NMF cho nguồn được làm bằng cách so với nguồn thật. Bộ tách mù thật phải tự gán, và đó là phần khó nhất. Báo con số tách nguồn phải nêu rõ mức thông tin mỗi phương pháp được cho.

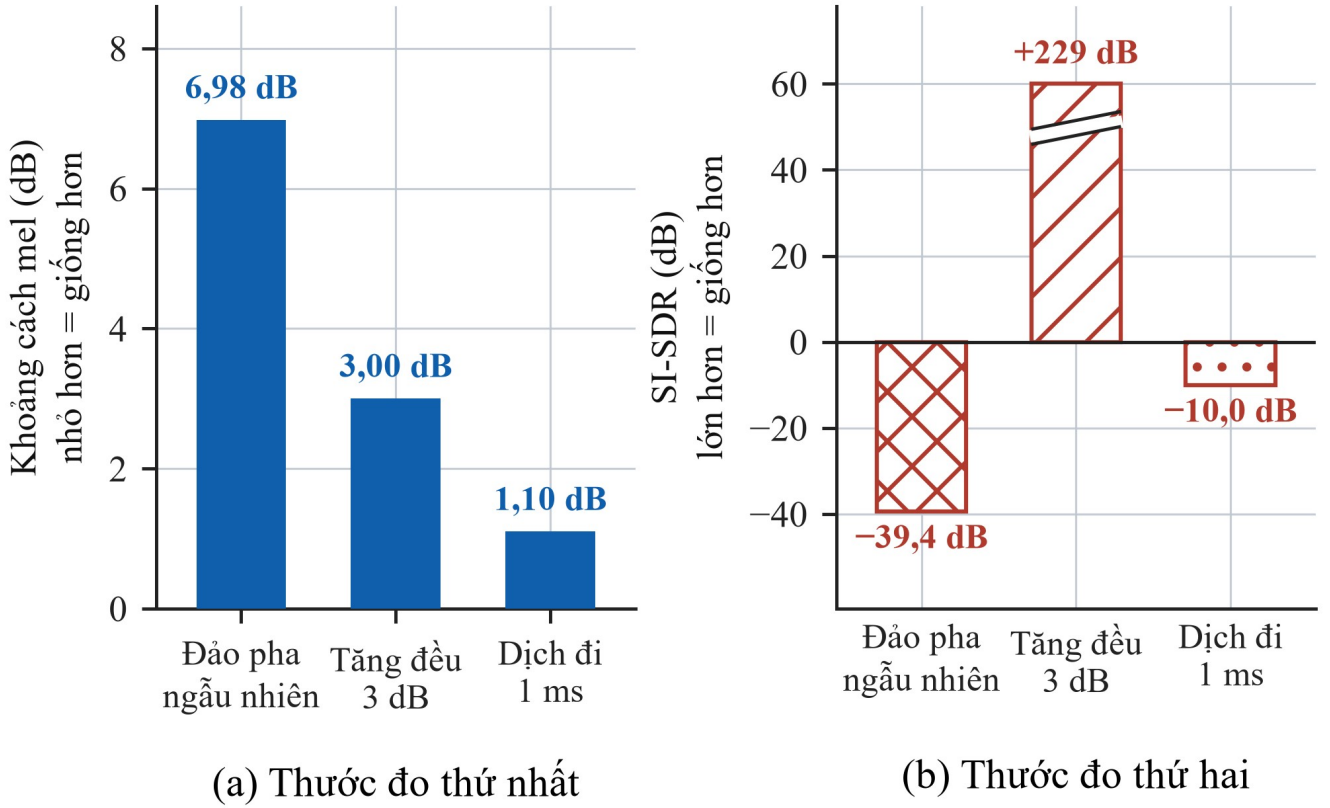
MỤC 13.8

Đánh giá — ba chỗ thước đo khách quan bị đánh lừa

- Mỗi thước đo trả lời một câu hỏi cụ thể.
- Khoảng cách mel và SI-SDR bất đồng ở cả ba phép biến đổi.
- Kết luận cuối cùng là phép nghe có đối chứng, với cỡ mẫu đủ.

Ba phép biến đổi, hai thước đo nói ngược nhau

- Đảo pha ngẫu nhiên, tăng đều 3 dB, dịch đi 1 ms.
- SI-SDR hỏi sóng có khớp không; khoảng cách mel hỏi phổ có khớp không.
- Không thước đo nào đúng cả ba lần.



Hình 13.26 — Ba phép biến đổi, và hai thước đo nói ngược nhau ở cả ba

Chọn thước đo theo câu hỏi

- Năm thước đo xếp theo câu hỏi mỗi thước trả lời.
- Cột thứ ba, thứ mà mỗi thước mù trước, quan trọng hơn hai cột đầu.
- Chọn sai thước đo cho kết quả đúng về một câu hỏi không được hỏi.

THƯỚC ĐO	TRẢ LỜI CÂU	MÙ TRƯỚC	DÙNG CHO
SI-SDR	Sóng có khớp không	Sai lệch tỉ lệ chung	Tách nguồn, khử nhiễu
Khoảng cách mel	Phổ có khớp không	Pha, và mọi thứ về pha	Vocoder, mô hình sinh phổ
Khoảng cách phổ log	Phổ khớp tới từng vạch	Pha; phạt mạnh vùng cao	Kiểm hồi quy của mục 12.8
LUFS và đỉnh	Có to đúng mức không	Toàn bộ nội dung	Chuẩn hoá đầu ra, mục 7.4
Nghe có đối chứng	Người có thích hơn không	Không mù gì cả	Kết luận cuối cùng và chỉ có nó

Nếu kết luận cuối cùng là phép nghe, thì phải nghe bao nhiêu người?

385	97	25	11
nếu hơn kém 5 %	nếu hơn kém 10 %	nếu hơn kém 20 %	nếu hơn kém 30 %

Cỡ mẫu tối thiểu, độ tin cậy 95 %. Nghe hai bản rồi kết luận là không đủ.

Hình 13.27 — Chọn thước đo theo câu hỏi, và cỡ mẫu cho phép nghe

Cỡ mẫu tối thiểu cho phép nghe so sánh cặp

Chênh lệch thật giữa hai bản	Số người nghe tối thiểu
Bản A hơn 5 % (55 % chọn A)	385
Bản A hơn 10 % (60 % chọn A)	97
Bản A hơn 20 % (70 % chọn A)	25
Bản A hơn 30 % (80 % chọn A)	11

Độ tin cậy 95 phần trăm. Chỉ có mười người nghe thì chọn hai bản khác nhau rõ rệt.

Bốn quy tắc

- Mù: người nghe không biết bản nào là bản nào.
- Ngẫu nhiên hoá thứ tự trình bày cho từng người nghe.
- Cùng mức to: chuẩn hoá LUFS (Loudness Units Full Scale – đơn vị độ to so với toàn thang) hai bản trước khi nghe (mục 7.4).
- Có mỏ neo: một bản chắc chắn kém, một bản chắc chắn tốt; ai không phân biệt được thì bỏ dữ liệu.

MỤC 13.9

Thực hành, bài tập và tổng kết

- Ba bài thực hành dùng chung mô-đun Python của học phần.
- Bảy bài tập mở rộng các phép đo của chương.
- Tổng kết bằng sáu con số.

Chạy xong trong vài giây

- Thực hành 1: dựng codec token bốn tầng; đo SNR khi giải mã một, hai và bốn tầng.
- Thực hành 2: tìm nhiệt độ thấp nhất mà tỉ lệ lặp chưa vượt 70 phần trăm.
- Thực hành 3: chạy lọc trung vị và NMF, so với mặt nạ tỉ lệ lý tưởng; nộp bảng bốn dòng.

Bài 1 đến bài 4

- Bài 1: lặp lại phép đo mục 13.3.3 với ngân sách 16 bit; cách chia tốt nhất có đổi không.
- Bài 2: chạy lại mục 13.3.4 với tiếng gõ và với giọng nói; đường cong có khác không.
- Bài 3: thêm mô hình bậc ba; dự đoán trước, đo, rồi so.
- Bài 4: thiết kế lịch nhiễu thứ ba giữa tuyến tính và cosine; tìm bước qua mốc 0 dB.

Chương 13 gói lại trong sáu con số

51×
nén được nhờ mã hoá thành token

66 %
tiết kiệm chỉ nhờ nhìn một token trước

496
bước: lịch cosine mới qua mốc 0 dB

64
bước lấy mẫu là đủ dùng, gần chạm sàn

34 dB
là trần của mọi phương pháp mặt nạ

−39 dB
khi giữ nguyên phổ mà đổi pha

Mô hình sinh ra cái gì thì dễ nói; cái nó không sinh ra mới quyết định tiếng nghe được.

Chương 14 mang những thứ này vào một sản phẩm thật.

Hình 13.28 — Chương 13 gói lại trong sáu con số

Ba điều đáng nhớ

Cái đi giữa các tầng quyết định nhiều hơn cái nằm trong mỗi tầng. Mọi bài toán đều có trần trên tính được; trần 34,06 dB của họ mặt nạ là ví dụ rõ nhất. Thước đo trả lời câu hỏi của chính nó, không trả lời câu hỏi của người dùng.

Những điều cần nhớ

1. Token bắt buộc vì số bước: ba phút nhạc là 7.938.000 mẫu nhưng chỉ 248.062 token.
2. Lượng tử hoá dư chồng mã sách nhỏ; hai đến bốn tầng là chỗ chọn tốt nhất, và bỏ tầng khi chạy là tính chất có sẵn.
3. Đo trên dữ liệu đã học là vô nghĩa; ngữ cảnh dài hơn cần dữ liệu nhiều hơn; nhiệt độ thấp gây suy sụp thành vòng lặp.
4. Khuếch tán: quá trình thuận biết trước; lịch cosine tốt hơn tuyến tính; sai lệch chạm sàn ở 64 đến 128 bước.
5. Phổ mel không có pha; Griffin-Lim giảm sai lệch độ lớn nhưng không đưa sóng về gần bản gốc.
6. Họ mặt nạ có trần 34,06 dB; thước đo phải chọn theo câu hỏi và phép nghe phải đủ cỡ mẫu.