Bài viết này đã được chấp nhận để xuất bản trong IEEE/ACM Transactions on Audio, Speech and Language Processing. Đây là phiên bản của tác giả chưa được biên tập đầy đủ và nội dung có thể thay đổi trước khi xuất bản cuối cùng. Thông tin trích dẫn: DOI 10.1109/TASLP.2023.3277693

Mã hóa lời nói thần kinh dự đoán miền tiềm ẩn

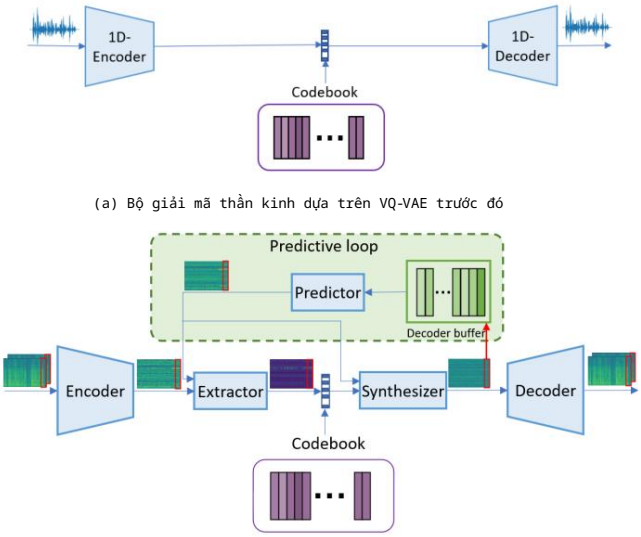
Xue Jiang, Xiulian Peng, Huaying Xue, Yuan Zhang, Yan Lu

Tóm tắt—Mã hóa giọng nói/âm thanh thần kinh gần đây đã chứng minh được khả năng cung cấp chất lượng cao ở tốc độ bit thấp hơn nhiều so với các phương pháp truyền thống. Tuy nhiên, các codec giọng nói/âm thanh thần kinh hiện có sử dụng các tính năng âm thanh hoặc các tính năng mù đã học được với mạng nơ-ron tích chập để mã hóa, theo đó vẫn có sự dư thừa về mặt thời gian trong các tính năng được mã hóa. Bài báo này giới thiệu mã hóa dự đoán miền tiềm ẩn vào khuôn khổ VQ-VAE để loại bỏ hoàn toàn các sự dư thừa như vậy và đề xuất TF-Codec để mã hóa giọng nói thần kinh có độ trễ thấp theo cách đầu cuối đến đầu cuối. Cụ thể, các tính năng được trích xuất được mã hóa theo điều kiện dự đoán từ các khung tiềm ẩn được lượng tử hóa trong quá khứ để các tương quan về mặt thời gian được loại bỏ thêm.
Hơn nữa, chúng tôi giới thiệu một nén có thể học được trên đầu vào tần số thời gian để điều chỉnh thích ứng sự chú ý dành cho các tần số chính và các chi tiết ở các tốc độ bit khác nhau. Một lược đồ lượng tử hóa vectơ có thể phân biệt được dựa trên ánh xạ khoảng cách đến mềm và Gumbel-Softmax được đề xuất để mô hình hóa tốt hơn các phân phối tiềm ẩn với ràng buộc tốc độ. Các kết quả chủ quan trên các tập dữ liệu giọng nói đa ngôn ngữ cho thấy, với độ trễ thấp, TF-Codec được đề xuất ở tốc độ 1 kbps đạt được chất lượng tốt hơn đáng kể so với Opus ở tốc độ 9 kbps và TF-Codec ở tốc độ 3 kbps vượt trội hơn cả EVS ở tốc độ 9,6 kbps và Opus ở tốc độ 12 kbps. Nhiều nghiên cứu được tiến hành để chứng minh hiệu quả của các kỹ thuật này.

Thuật ngữ chỉ mục—Mã hóa âm thanh/lời nói thần kinh, mã hóa tự động, mã hóa dự đoán.

I. GIỚI THIỆU

Trong những năm gần đây, mã hóa giọng nói/âm thanh thần kinh đã phát triển nhanh chóng và hiện cung cấp kết quả chất lượng cao ở tốc độ bit rất thấp, đặc biệt là đối với giọng nói. Các codec thần kinh hiện có chủ yếu có thể được chia thành hai loại, codec dựa trên mô hình giải mã tạo sinh [1]-[5] và mã hóa giọng nói/âm thanh thần kinh đầu cuối [6]-[12]. Loại trước trích xuất các đặc điểm âm thanh từ âm thanh để mã hóa và sử dụng bộ giải mã mạnh mẽ để tái tạo dạng sóng dựa trên các mô hình tạo sinh. Loại sau chủ yếu tận dụng khuôn khổ VQ-VAE [13] để học bộ mã hóa, bộ lượng tử hóa vectơ và bộ giải mã theo cách đầu cuối. Các đặc điểm tiềm ẩn cần được lượng tử hóa chủ yếu được học một cách mù quáng bằng cách sử dụng mạng nơ-ron tích chập (CNN) mà không có bất kỳ kiến thức nào trước đó. Các phương pháp này đã cải thiện đáng kể hiệu quả mã hóa bằng cách đạt được chất lượng cao ở tốc độ bit thấp. Tuy nhiên, các tương quan thời gian không được khai thác đầy đủ trong các thuật toán này, dẫn đến nhiều sự trùng lặp giữa các khung lân cận trong các đặc điểm được mã hóa. Theo quan điểm này,



(a) Bộ giải mã thần kinh dựa trên VQ-VAE trước đó

(b) Bộ giải mã thần kinh dựa trên VQ-VAE dự đoán của chúng tôi

Hình 1. Đề xuất mã hóa giọng nói thần kinh dự đoán miền tiềm ẩn.

chúng tôi đề xuất kết hợp mã hóa dự đoán vào khuôn khổ mã hóa thần kinh dựa trên VQ-VAE để loại bỏ những sự dư thừa như vậy.

Mã hóa dự đoán được sử dụng rộng rãi trong hình ảnh truyền thống [14], video [15]-[17] và mã hóa âm thanh [18], [19] để loại bỏ sự dư thừa về không gian và thời gian, trong đó các khối/khung/mẫu lân cận được tái tạo được sử dụng để dự đoán khối/khung/mẫu hiện tại và các phần dư được dự đoán được lượng tử hóa và mã hóa thành một luồng bit. Các phần dư sau khi dự đoán thưa hơn nhiều và entropy của chúng giảm đáng kể. Trong codec video thần kinh [20], [21], mối tương quan thời gian như vậy cũng được khai thác bằng cách sử dụng khung tham chiếu được căn chỉnh theo chuyển động làm dự đoán hoặc ngữ cảnh để mã hóa khung hiện tại. Tuy nhiên, trong codec âm thanh thần kinh, các kỹ thuật như vậy hiếm khi được nghiên cứu, theo hiểu biết của chúng tôi.

Mặc dù các tương quan thời gian được khai thác trong bộ mã hóa và giải mã mã hóa âm thanh/lời nói của mạng nơ-ron tích chập hoặc hồi quy, các hoạt động này có thể được coi là một loại dự đoán vòng hở hoặc biến đổi phi tuyến tính (Xem Hình 1 (a)). Sau khi lượng tử hóa, tương quan thời gian ở phía bộ giải mã bị phá vỡ ở một mức độ nào đó. Chúng tôi thấy rằng để phục hồi tốt hơn từ tiếng ồn lượng tử hóa ở tốc độ bit thấp, mạng nơ-ron có xu hướng bảo toàn một số sự dư thừa trong biểu diễn tiềm ẩn đã học. Tuy nhiên, bằng cách sử dụng dự đoán vòng kín như trong mã hóa dự đoán của chúng tôi (xem Hình 1 (b)), sự dư thừa như vậy bị loại bỏ trong các tính năng được mã hóa nhưng khả năng phục hồi không bị ảnh hưởng đối với dự đoán vòng kín. Các tính năng tiềm ẩn đã học rất thưa thớt và quá trình giải mã có thể đạt được khả năng phục hồi chất lượng cao bằng cách sử dụng cùng một dự đoán được sử dụng trong quá trình mã hóa.

Xue Jiang đang làm việc tại Khoa Kỹ thuật Thông tin và Truyền thông, Đại học Truyền thông Trung Quốc, Bắc Kinh 100024, Trung Quốc (email: jiangxhoho@cuc.edu.cn).

X. Peng, H. Xue và Y. Lu làm việc tại Microsoft Research Asia, Bắc Kinh 100080, Trung Quốc (e-mail: xipe@microsoft.com; huxue@microsoft.com; yanlu@microsoft.com).

Y. Zhang đang làm việc tại Phòng thí nghiệm trọng điểm nhà nước về hội tụ phương tiện và truyền thông, Đại học Truyền thông Trung Quốc, Bắc Kinh 100024, Trung Quốc (email: yzhang@cuc.edu.cn).

Tác giả liên hệ là Yan Lu và Xiulian Peng.

Công trình này được thực hiện khi Xue Jiang còn là thực tập sinh tại Microsoft Research Asia.

Bài viết này đã được chấp nhận để xuất bản trong IEEE/ACM Transactions on Audio, Speech and Language Processing. Đây là phiên bản của tác giả chưa được biên tập đầy đủ và nội dung có thể thay đổi trước khi xuất bản cuối cùng. Thông tin trích dẫn: DOI 10.1109/TASLP.2023.3277693

Bài báo này là bài báo đầu tiên giới thiệu mã hóa dự đoán vào khuôn khổ VQ-VAE để mã hóa giọng nói thần kinh. Để giảm độ trễ, mã hóa dự đoán này được thực hiện trong miền tiềm ẩn như thể hiện trong Hình 1 (b). Không giống như mã hóa video/âm thanh dự đoán truyền thống, trừ các mẫu khỏi dự đoán, chúng tôi giới thiệu một trình trích xuất có thể học được để hợp nhất dự đoán với các tính năng mã hóa, thu được thông tin "mới" thưa thớt để mã hóa từng khung. Tất cả các mô-đun đều được học từ đầu đến cuối với đào tạo đối nghịch. Hơn nữa, không giống như hầu hết các codec thần kinh trước đây sử dụng đầu vào miền thời gian, chúng tôi giới thiệu đầu vào tần số thời gian với nén có thể học được trên biên độ. Điều này cho phép mạng tự động cân bằng sự chú ý dành cho các thành phần chính và chi tiết ở các tốc độ bit khác nhau (xem Hình 1 (b)), tăng đáng kể chất lượng ở tốc độ bit thấp tới 1 kbps cho mã hóa giọng nói có độ trễ thấp.

Những đóng góp chính của bài báo này được tóm tắt như sau:

- Chúng tôi đề xuất TF-Codec, một bộ giải mã giọng nói thần kinh có độ trễ thấp, theo hiểu biết của chúng tôi, là bộ giải mã thời gian thực đầu tiên báo cáo chất lượng cao ở mức 1 kbps.
- Chúng tôi giới thiệu mã hóa dự đoán vào bộ giải mã giọng nói thần kinh dựa trên VQ-VAE, giúp giảm đáng kể sự dư thừa về mặt thời gian và do đó tăng hiệu quả mã hóa. • Chúng tôi giới thiệu một phương pháp nén có thể học được trên đầu vào tần số thời gian để điều chỉnh thích ứng sự chú ý dành cho các thành phần chính và chi tiết ở các tốc độ bit khác nhau. • Chúng tôi giới thiệu một cơ chế lượng tử hóa vectơ có thể phân biệt được dựa trên ánh xạ khoảng cách đến mềm và Gumbel-Softmax để tạo điều kiện kiểm soát tốc độ và đạt được tối ưu hóa tốc độ-biến dạng tốt hơn. • Chúng tôi thảo luận về các cách để tăng cường tính mạnh mẽ của mã hóa giọng nói thần kinh dự đoán trong điều kiện mất gói tin, đã đạt được những kết quả đầy hứa hẹn.

II. CÔNG TRÌNH LIÊN QUAN

A. Mã hóa giọng nói/âm thanh thần kinh Mã

hóa âm thanh dựa trên mô hình tạo sinh Với sự tiến bộ của các mô hình tạo sinh trong việc cung cấp tổng hợp giọng nói chất lượng cao, các nhà nghiên cứu gần đây đã đề xuất tận dụng chúng cho mã hóa giọng nói [1]–[5], chẳng hạn như WaveNet [22] và LPCNet [23]. WaveNet là mô hình đầu tiên được sử dụng làm bộ giải mã tạo sinh học để tạo ra âm thanh chất lượng cao từ bộ mã hóa thông thường ở tốc độ 2,4 kbps [1]. Một số nhà nghiên cứu [5] đã cải thiện chất lượng giọng nói Opus ở tốc độ bit thấp bằng cách sử dụng LPCNet để tổng hợp giọng nói. Lyra [2] là một mô hình tạo sinh tổng hợp giọng nói từ phổ mel lượng tử hóa bằng mô hình WaveGRU tự hồi quy, tạo ra giọng nói chất lượng cao ở tốc độ 3 kbps. Mặc dù các phương pháp này đạt được chất lượng tốt ở tốc độ bit thấp, nhưng tiềm năng đầy đủ của mã hóa âm thanh thần kinh vẫn chưa được khai thác.

Mã hóa âm thanh đầu cuối Thể loại này học mã hóa en, lượng tử hóa vectơ và giải mã theo cách đầu cuối dựa trên khuôn khổ VQ-VAE [13] [6]–[12]. Trong [6], bộ mã hóa VQ-VAE và bộ giải mã dựa trên WaveNet được học chung từ đầu đến cuối, mang lại chất lượng tái tạo cao trong khi truyền giọng nói qua biểu diễn tiềm ẩn nhỏ gọn tương ứng với tốc độ bit rất thấp. Đề xuất gần đây SoundStream [8] đạt được chất lượng âm thanh vượt trội ở nhiều tốc độ bit từ 3 kbps đến 18 kbps với khả năng học từ đầu đến cuối và kết hợp giữa mất mát đối nghịch và tái tạo.

Gần đây hơn, một codec âm thanh đầu cuối với đường ống mã hóa dư thừa mô-đun chéo đã được đề xuất cho mã hóa có thể mở rộng [10]. Không giống như các phương pháp trước đây dựa trên đầu vào dạng sóng với phép tích chập 1D, TFNet [12] gần đây lấy đầu vào tần số thời gian với mô hình bộ mã hóa-bộ giải mã-lọc thời gian 2D nhân quả cho mã hóa giọng nói đầu cuối. Trong số tất cả các phương pháp này, các tính năng tiềm ẩn từ bộ mã hóa chủ yếu được học một cách mù quáng mà không có bất kỳ thông tin trước nào và thường vẫn còn các tương quan thời gian trong chúng. Trong bài báo này, chúng tôi đề xuất mã hóa dự đoán để loại bỏ thêm các phần dư thừa.

B. Mã hóa dự đoán

Nén âm thanh cổ điển Mã hóa dự đoán được sử dụng rộng rãi trong mã hóa âm thanh cổ điển [18], [19], [24]. Vì các mẫu âm thanh liên tiếp có mối tương quan cao, thay vì lượng tử hóa và truyền các mẫu âm thanh độc lập, phần dư giữa mẫu hiện tại và dự đoán của nó dựa trên các mẫu trước đó được mã hóa. DPCM [18]/ ADPCM [19] thường sử dụng dự đoán ngược (còn được gọi là dự đoán vòng kín) trong đó các mẫu được tái tạo trong quá khứ được sử dụng để có được dự đoán, có thể có một số điều chỉnh cho bộ dự đoán và bộ lượng tử trong ADPCM. Một kỹ thuật được sử dụng rộng rãi khác trong mã hóa và xử lý giọng nói, mã hóa dự đoán tuyến tính (LPC) [24], tận dụng bộ dự đoán tuyến tính để ước tính các mẫu trong tương lai dựa trên mô hình bộ lọc nguồn. Các hệ số bộ lọc tuyến tính trong LPC được tính theo cách vòng hở với dự đoán thuận, trong đó các mẫu gốc thay vì các mẫu được tái tạo, được sử dụng để dự đoán. Các phần dư, cùng với các hệ số tuyến tính, được lượng tử hóa và mã hóa.

Nén video/hình ảnh cổ điển Các tiêu chuẩn mã hóa video truyền thống [15]–[17] luôn áp dụng mô hình mã hóa dự đoán để loại bỏ sự dư thừa về mặt thời gian, trong đó dự đoán được tạo ra bằng cách ước tính và bù chuyển động dựa trên khối, và phần còn lại giữa khung gốc và dự đoán được chuyển đổi, lượng tử hóa và mã hóa entropy.

Trong mã hóa hình ảnh [14] và mã hóa trong khung của video [15]– [17], các khối lân cận được tái tạo được sử dụng để dự đoán khối hiện tại, trong miền tần số hoặc pixel, và các phần dư được dự đoán được mã hóa.

Nén video sâu Trong mã hóa video nơ-ron, một cách tiếp cận điển hình là thay thế các mô-đun thủ công, chẳng hạn như ước tính chuyển động, bằng mạng nơ-ron, trong khi vẫn tuân thủ mô hình mã hóa dự đoán. DVC [20] cung cấp các dự đoán thời gian chính xác hơn bằng cách đào tạo chung các mạng ước tính chuyển động và bù trừ. Thông tin còn lại sau khi dự đoán sau đó được mã hóa bởi mạng mã hóa còn lại.

Công trình có liên quan nhất, DCVC [21], thay vào đó đề xuất một sự thay đổi mô hình từ mã hóa dự đoán sang mã hóa có điều kiện. Nó giới thiệu thông tin ngữ cảnh thời gian phong phú như một điều kiện cho cả bộ mã hóa và bộ giải mã, cải thiện đáng kể hiệu quả mã hóa.

Mã hóa giọng nói thần kinh dự đoán Có một số nỗ lực trong lĩnh vực công việc này có liên quan nhất đến công việc của chúng tôi. Công việc đồng thời [25] giới thiệu mã hóa dự đoán

Bài viết này đã được chấp nhận để xuất bản trong IEEE/ACM Transactions on Audio, Speech and Language Processing. Đây là phiên bản của tác giả chưa được biên tập đầy đủ và nội dung có thể thay đổi trước khi xuất bản cuối cùng. Thông tin trích dẫn: DOI 10.1109/TASLP.2023.3277693

trong miền tham số, trong đó một bộ dự đoán dựa trên đơn vị hồi quy có cổng (GRU) được áp dụng để dự đoán các hệ số LPC từ quá khứ. Tuy nhiên, vì nó dựa trên phân tích LPC theo giả định bộ lọc nguồn, nên tiềm năng của nó không được khai thác đầy đủ và không thể dễ dàng mở rộng sang các miền tín hiệu khác như âm nhạc. Một công trình có liên quan khác, mã hóa nhận thức hai giai đoạn của giọng nói [26], tận dụng mã hóa dự đoán trong không gian tiềm ẩn như một chiến lược học biểu diễn. Khác với động cơ của chúng tôi, nó tập trung vào việc học biểu diễn mà không coi dự đoán là một mô-đun mã hóa.

Dựa trên các phương pháp này, chúng tôi đưa mã hóa dự đoán vào khuôn khổ VQ-VAE để mã hóa âm thanh thần kinh, để loại bỏ tốt hơn các dư thừa về mặt thời gian và đạt được hiệu quả mã hóa tốt hơn. Không giống như [25], mã hóa dự đoán hoạt động trong miền tiềm ẩn và được đào tạo với các mô-đun VQ-VAE khác từ đầu đến cuối, do đó khám phá sâu rộng tiềm năng của nó.

C. Mô hình hồi quy tự động

Các mô hình tạo tự hồi quy đã chứng minh khả năng mạnh mẽ trong tổng hợp giọng nói [22], [27]. Chúng thường tạo các mẫu âm thanh từng cái một theo cách tự hồi quy, trong đó các mẫu được tạo trước đó được sử dụng để tạo mẫu hiện tại. Mã hóa dự đoán của chúng tôi cũng sử dụng phương pháp tự hồi quy, nhưng trái ngược với tự hồi quy miền mẫu, nó hoạt động trong miền tiềm ẩn để giảm độ trễ của vòng lặp tự hồi quy. Vòng lặp này chỉ vượt qua lớp lượng tử hóa và không yêu cầu phải đi qua bộ giải mã để có được đầu ra cho tự hồi quy.

D. Lượng tử hóa vectơ

Lượng tử hóa vectơ (VQ) là một kỹ thuật cơ bản được sử dụng rộng rãi trong các codec âm thanh truyền thống như Opus [28] và CELP [29]. Gần đây, nó cũng đã được áp dụng cho việc học biểu diễn rời rạc [13] và đóng vai trò là cơ sở của mã hóa âm thanh thần kinh đầu cuối [6]–[12]. Vì lượng tử hóa vốn không thể phân biệt được, nên một số phương pháp đã được đề xuất trong tài liệu để cho phép học đầu cuối trong mã hóa âm thanh thần kinh, bao gồm phương pháp có mất cam kết trong VQ-VAE [13], EMA [13], phương pháp dựa trên Gumbel-Softmax [30] [31] và kỹ thuật mềm sang cứng [32]. VQ-VAE [13] xấp xỉ đạo hàm bằng hàm danh tính sao chép trực tiếp các gradient từ đầu vào bộ giải mã sang đầu ra bộ mã hóa. Sổ mã được học bằng cách di chuyển từ mã đã chọn, được tìm thấy thông qua một số phép đo khoảng cách, về phía các tính năng của bộ mã hóa.

Ngược lại, các phương pháp Gumbel-Softmax và soft-to-hard đưa xác suất chọn một từ mã vào VQ, cho phép chọn các từ mã rời rạc theo cách có thể phân biệt được. Tuy nhiên, phương pháp trước sử dụng phép chiếu tuyến tính với Gumbel-Softmax để có được xác suất, phương pháp này thiếu mối tương quan rõ ràng với lỗi lượng tử hóa. Phương pháp sau ánh xạ khoảng cách thành xác suất và sử dụng phép gán mềm với quá trình ủ trong quá trình đào tạo, có khả năng dẫn đến khoảng cách giữa quá trình đào tạo và suy luận. Được thúc đẩy bởi các công trình này, chúng tôi đề xuất một lược đồ Distance-Gumbel-Softmax với điều khiển tốc độ ánh xạ rõ ràng lỗi lượng tử thành xác suất trong khi sử dụng phép gán cứng trong quá trình đào tạo và suy luận.

III. KẾ HOẠCH ĐỀ XUẤT

A. Tổng quan

Hãy để x biểu thị tín hiệu miền thời gian cần được mã hóa và x^ là tín hiệu được phục hồi sau khi giải mã. Việc tối ưu hóa mã hóa âm thanh thần kinh nhằm mục đích giảm thiểu độ méo tín hiệu được phục hồi Dist(x, x^|θ) theo một ràng buộc tốc độ nhất định, tức là R(x|θ) ≤ Rtarget. θ biểu thị các tham số mạng nơ-ron.

Trong bài báo này, chúng tôi tập trung vào mã hóa giọng nói có độ trễ thấp.

Như thể hiện trong Hình 2, chúng tôi sử dụng một bộ mã hóa để trích xuất tiềm ẩn biểu diễn $x_R = \{x_{R0}, x_1, \ldots, x_{RT}\}$ từ x, trong đó x_{Rt} là đặc điểm giọng nói tại khung t và T là tổng số của khung hình. Đối với mỗi khung hình x_{Rt} trong x_R, một dự đoán x_{Pt} học được từ các mã tiềm ẩn được tái tạo trong quá khứ $\hat{x}_R$ thông qua một bộ dự đoán fpred với trường tiếp nhận của N khung quá khứ, được đưa ra bởi $x_{Pt} = f_{pred}(\hat{x}_{R,t-N}, \ldots, \hat{x}_{R,t-1})$. Dự đoán này đóng vai trò là một bối cảnh thời gian cho cả mã hóa và giải mã. Đối với mã hóa, một bộ trích xuất fextr học thông tin giống như phần dư theo x_t), là "mới" đối với các khung quá khứ. x_{Nt} từ cả hai x_{Rt} và x_{Pt} Với hoạt động tự hồi quy này, sự dư thừa về mặt thời gian có thể được giảm hiệu quả. Đặc điểm giống như phần dư được trích xuất $x_N = \{x_{N0}, x_1, \ldots, x_{NT}\}$ là sau đó được lượng tử hóa thông qua một sổ mã được học bởi Distance-Gumbel-Softmax và được mã hóa entropy bằng mã hóa Huffman. trong $\hat{x}_N$ Để giải mã, giống như dư lượng hóa $\hat{x}$ được hợp nhất với dự đoán x_{Pt} fsynr để đặc điểm có được mã tiềm ẩn tái tạo hiện tại $\hat{x}_{Rt}$ thông qua một bộ tổng hợp theo $\hat{x}_R$ = fsynr($\hat{x}$ tái tạo dạng sóng $\hat{x}$. Chúng tôi áp dụng đào t , được tạo đối x_t nghịch để đạt x_{Nt}, x_{Pt} đưa ra). Sau đó, một bộ giải mã được sử dụng để được chất lượng nhận thức tốt. Trong các tiểu mục sau, chúng tôi sẽ mô tả chi tiết các kỹ thuật này.

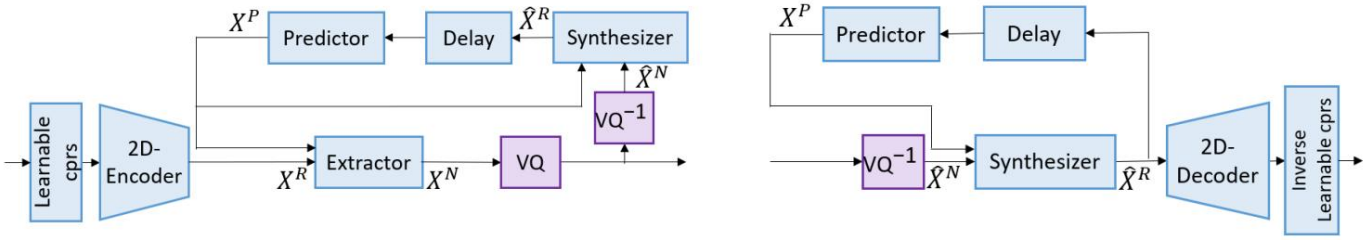
B. Nén đầu vào có thể học được

Dạng sóng đầu vào x đầu tiên được chuyển đổi thành miền tần số với phép biến đổi fourier thời gian ngắn (STFT), tạo ra phổ thời gian-tần số X_R trong đó T là số khung, F là số bin tần số và 2 biểu thị phần ảo $\in \mathbb{R}^{2\times F\times T}$ và phần thực của phổ phức hợp. Chúng tôi lấy đầu vào miền tần số thay vì miền thời gian được áp dụng rộng rãi trong các công trình trước đây [8], [13], vì miền tần số phù hợp tốt với nhận thức của con người.

Trong miền này, một số đặc điểm của giọng nói (như tần số cơ bản, sóng hài và sóng formant) được thể hiện rõ ràng, giúp bộ mã hóa dễ dàng học các tính năng liên quan đến nhận thức của con người để mã hóa. Vì đầu vào miền tần số thường thể hiện dải động cao và phân phối cực kỳ mất cân bằng do sóng hài, chúng tôi sử dụng nén theo luật lũy thừa từng phần tử trên phần biên độ được chỉ định bởi |X| p trong đó |X| là phổ biên độ của X, trong khi pha được giữ nguyên, tạo ra phổ thời gian-tần số nén $X_{cprs} \in \mathbb{R}^{2\times F\times T}$ Nén hoạt động như một dạng chuẩn hóa đầu vào, cân bằng tầm quan trọng của các tần số khác nhau và đảm bảo quá trình đào tạo ổn định hơn.

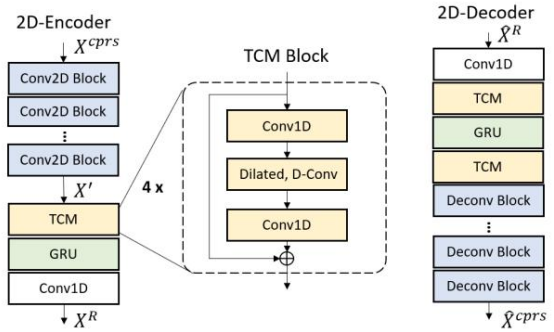
Hơn nữa, chúng tôi làm cho tham số công suất p có thể học được trong quá trình đào tạo, cho phép mô hình thích ứng với các tốc độ bit khác nhau. Cụ thể, ở tốc độ bit thấp, p cao hơn có thể được ưu tiên vì nó dẫn đến sự chú ý nhiều hơn đến các thành phần chính, trong khi ở tốc độ bit cao, có thể chú ý nhiều hơn đến các chi tiết với tốc độ bit thấp hơn

Bài viết này đã được chấp nhận để xuất bản trong IEEE/ACM Transactions on Audio, Speech and Language Processing. Đây là phiên bản của tác giả chưa được biên tập đầy đủ và nội dung có thể thay đổi trước khi xuất bản cuối cùng. Thông tin trích dẫn: DOI 10.1109/TASLP.2023.3277693



(a) Mã hóa (b) Giải mã

Hình 2. Các mô-đun mã hóa và giải mã cho phương pháp đề xuất.



Hình 3. Kiến trúc của bộ mã hóa và bộ giải mã. D-Conv biểu thị phép tích chập theo chiều sâu.

p. Kỹ thuật này đã được chứng minh là đặc biệt hiệu quả đối với mã hóa tốc độ bit rất thấp, chẳng hạn như 1 kbps, trong các thí nghiệm của chúng tôi.

C. Bộ mã hóa và giải mã

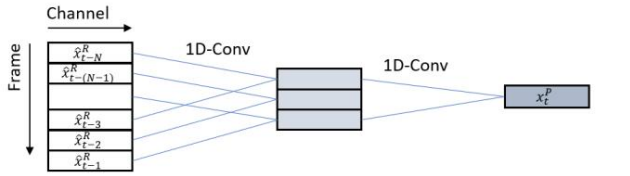
Bộ mã hóa lấy phổ tần số thời gian nén Xcprs R làm đầu vào. Như thể hiện trong Hình 3, 2×F×T năm lớp tích chập nhân quả 2D đầu tiên được sử dụng để giải tương quan của nó trong hai chiều (F, T) với kích thước hạt nhân là (5, 2), các kênh đầu ra là 16, 32, 32, 64 và 64, và bước tiến là 1, 2, 4, 4 và 2 dọc theo chiều tần số F. Chiều thời gian T được giữ nguyên mà không lấy mẫu lại. Chiều đầu ra sau năm lớp tích chập là C × F × T, sau đó chúng tôi gấp tất cả thông tin tần số vào chiều kênh, tạo ra X′ RC = C′ Để nắm bắt các phụ thuộc thời gian tầm xa, chúng tôi sử dụng thêm một mô-đun tích chập thời gian (TCM) [33] với các tích chập theo chiều sâu giãn nở nhân quả theo sau là một khối đơn vị hồi quy có cổng (GRU) trên C×T như trong × F′ .

[12], để nắm bắt cả các phụ thuộc thời gian ngắn hạn và dài hạn. Lớp tích chập 1D cuối cùng với kích thước hạt nhân là 1 được sử dụng để thay đổi chiều kênh thành Cd để lượng tử hóa bằng mã hóa dự đoán. Bộ mã hóa cuối cùng tạo ra đầu ra XR = {x Bộ giải mã ngược lại với bộ mã hóa, tái tạo xˆ R } Để phục hồi tốt hơn, nhiều mô-đun TCM được sử dụng trong bộ giải mã hơn là trong bộ mã hóa. Cụ thể, một mô-đun TCM, một khối GRU và một mô-đun TCM khác được sử dụng theo cách xen kẽ để nắm bắt các phụ thuộc thời gian cục bộ và toàn cục ở các độ sâu khác nhau. Giải tích nhân quả được sử dụng để khôi phục độ phân giải tần số thành

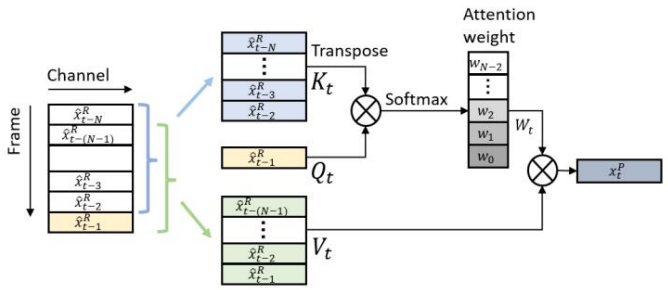
F Tỷ lệ cược 0 , 1 , ... Rx T } R Cd×T .

và

xˆ từ các đặc điểm được tái tạo Xˆ R = {xˆ R 0 , xˆ 1 R , ... , T R Cd×T .



(a) Bộ dự đoán dựa trên tích chập



(b) Bộ dự báo thích ứng

Hình 4. Cấu trúc mạng của bộ dự báo.

bộ giải mã đưa ra một tính năng Xˆ cprs R Sau khi nén biên độ nghịch đảo và 2×F×T STFT nghịch đảo, dạng sóng xˆ cuối cùng được tái tạo. Toàn bộ quá trình là nhân quả để có thể đạt được độ trễ thấp.

D. Mã hóa dự đoán miền tiềm ẩn

Vì mã hóa dự đoán là tự hồi quy, để giảm độ trễ, chúng tôi chỉ nghiên cứu nó trong miền tiềm ẩn, như thể hiện trong Hình 1 (b) và sự phân tách mã hóa/giải mã trong Hình 2.

Bộ dự đoán Bộ dự đoán cung cấp dự đoán về khung hiện tại từ quá khứ, được đưa ra bởi i = 1, 2, ..., N) R Cd với một cửa sổ là

x P t = fpred(xˆ R t-tôi ,

N khung. Như thể hiện trong Hình 4, chúng tôi nghiên cứu hai phương pháp cho dự đoán này: (1) Bộ dự đoán dựa trên tích chập, sử dụng hai lớp tích chập 1D được kích hoạt bởi ReLU tham số (PReLU) [34] để đạt được trường tiếp nhận là 280 ms.
(2) Bộ dự đoán thích ứng, học hạt nhân dự đoán từ quá khứ để thích ứng với tín hiệu giọng nói thay đổi theo thời gian. Hạt nhân được suy ra từ quá khứ dựa trên giả định rằng các hệ số dự đoán tuyến tính là hằng số cục bộ.
Cụ thể, nó sử dụng một cơ chế tương tự như sự chú ý của bản thân [35] với truy vấn Qt, khóa Kt và ma trận giá trị Vt,

Bài viết này đã được chấp nhận để xuất bản trong IEEE/ACM Transactions on Audio, Speech and Language Processing. Đây là phiên bản của tác giả chưa được biên tập đầy đủ và nội dung có thể thay đổi trước khi xuất bản cuối cùng. Thông tin trích dẫn: DOI 10.1109/TASLP.2023.3277693

được định nghĩa như sau

$$Qt = [x\hat{}_t^R 1]^T \quad R1 \times Cd$$
$$Kt = [x\hat{}_{t\,2}^R, Rx\hat{}_{t\,3}^R, ..., x\hat{}_{t\,N}^R]^T \quad R(N\ 1)\times Cd \qquad (1)$$
$$Vt = [x\hat{}_{t\,2}^R, RRx\hat{}_{t\,2}^R x\hat{}_{t\,3}^R, ..., x\hat{}_{t(N\ 1)}^R]^T \quad R(N\ 1)\times Cd$$

Nó học một ma trận trọng số chú ý Wt R 1×(N 1) đóng vai trò là hạt nhân dự đoán bằng

$$Wt = Sof\,tmax(Qt \cdot (Kt)^T / \sqrt{Cd}), \qquad (2)$$

trong đó Sof tmax(·) là hàm softmax. Ma trận trọng số chú ý sau đó được nhân với Vt để có được dự đoán R Cd×1 theo [x chú ý trong cách nó nắm số chú $x^P_{t\,] = (Trọng\ lượng \cdot Vt)}$ bất trọng . Phương pháp này tương tự như tự ý một cách thích ứng từ các tính năng đầu vào, nhưng ở đây chúng tôi mở rộng nó như một loại dự đoán.

Chúng tôi sẽ trình bày sự so sánh giữa hai loại yếu tố dự báo này trong phần thực nghiệm.

Để hướng dẫn bộ dự báo đưa ra dự đoán thời gian tốt để loại bỏ sự dư thừa, một tổn thất dự đoán được đưa vào trong quá trình đào tạo như

$$Lpred = Ví\ dụ[\frac{1}{CdT}\sum_{t=1}^{T} ||x^P_t, sg(x^R_t)||_2^2], \qquad (3)$$

trong đó sg(·) là toán tử stop-gradient. Ở giai đoạn đào tạo ban đầu, khi cả bộ mã hóa và bộ dự đoán đều có xu hướng hoạt động kém, việc ép buộc Lpred x t có thể khiến bộ mã hóa bị nhầm lẫn về việc tạo ra các biểu diễn tốt hơn để tái tạo. Do đó, chúng tôi thêm toán tử stop-gradient vào đầu ra của bộ mã hóa x để đào tạo ổn định hơn và tái tạo tốt hơn.

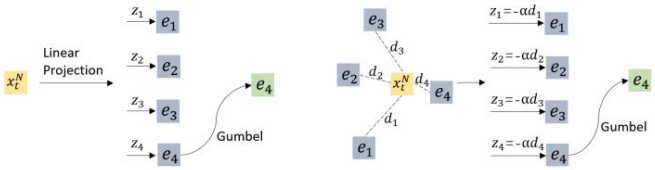
x^R_t

Bộ trích xuất và bộ tổng hợp Cả bộ trích xuất fextr và bộ tổng hợp fsynr đều bao gồm một lớp tích chập 1D với kích thước hạt nhân là 1 và bước tiến là 1, theo sau là chuẩn hóa theo lô (BN) và ReLU tham số làm hàm kích hoạt phi tuyến tính.

Thuật toán đào tạo Như đã thảo luận trước đó, vòng lặp dự đoán hoạt động theo cách tự hồi quy trong miền tiềm ẩn, vì bộ dự đoán cần bối cảnh được tái tạo trong Xˆ R để dự đoán. Không dễ để sử dụng ép buộc của giáo viên cho hồi quy tự động miền tiềm ẩn vì nhiễu lượng tử hóa vectơ không dễ mô hình hóa trong quá trình tối ưu hóa đầu cuối. Thay vào đó, chúng tôi áp dụng một chiến lược tạo điều kiện cho đào tạo song song để cải thiện hiệu quả đào tạo. Cụ thể, tại mỗi lần lặp, chúng tôi tận dụng mô hình từ lần lặp trước để có được Xˆ R để dự đoán trong lần lặp hiện tại. Sau đó, vòng lặp dự đoán có thể được đào tạo song song trong lần lặp hiện tại, do đó tăng tốc quá trình đào tạo.

E. Lượng tử hóa vectơ với điều khiển tốc độ

Như đã thảo luận trong Phần II.D, các phương pháp Gumbel-Softmax [30] [31] và mềm-đến-cứng [32] đưa ra xác suất chọn một từ mã, do đó làm cho việc kiểm soát tốc độ trở nên khả thi. Tuy nhiên, Gumbel-Softmax sử dụng phép chiếu tuyến tính để chọn từ mã mà không liên hệ rõ ràng với lỗi lượng tử hóa, như minh họa trong Hình 5 (a). Hoạt động trung bình có trọng số trong phương pháp mềm-đến-cứng có thể dễ dàng dẫn đến khoảng cách giữa đào tạo và suy luận. Theo quan điểm này



(a) Gumbel-Softmax (b) Khoảng cách-Gumbel-Softmax

Hình 5. Cơ chế lượng tử hóa vectơ. (a) Gumbel-Softmax trong [30]. xN tiềm ẩn được chiếu tới logit zi thông qua phép chiếu tuyến tính và chuyển thành xác suất với Gumbel-Softmax. (b) Khoảng cách-Gumbel-Softmax của chúng tôi. Khoảng cách giữa xN tiềm ẩn và các từ mã ei đầu tiên được tính toán và sau đó được ánh xạ tới xác suất với Gumbel-Softmax.

hạn chế, chúng tôi đề xuất phương pháp Distance-Gumbel-Softmax, như thể hiện trong Hình 5 (b), để lượng tử hóa, tận dụng lợi thế của cả hai phương pháp để cung cấp phép gán có nhận biết lỗi lượng tử hóa với khả năng kiểm soát tốc độ.

VQ dựa trên Distance-Gumbel-Softmax Như thể hiện trong Hình 5 (b), cho một sổ mã có K từ mã E = {e1, e2, ..., eK} R Cd×K, trước tiên chúng ta tính toán khoảng cách R Cd và tất cả K giữa vectơ tiềm ẩn hiện tại x mã từ như x^N_t

$$dt = [\ell d(x^N_t, e1), ..., \ell d(x^N_t, eK)] \quad R\ K, \qquad (4)$$

trong đó ℓd là một phép đo khoảng cách và chúng tôi sử dụng ℓ2 trong quá trình triển khai của mình. Sau đó, khoảng cách được ánh xạ tới logit zt, được chỉ định bởi zt = α · dt, trong đó α là một số vô hướng dương để kiểm soát việc ánh xạ từ khoảng cách $\ell d(x^N_t, ek)$ tới logit zt,k sao gần hơn với tính năng hiện tại x cho từ mã sẽ có xác suất được chọn cao hơn và chúng tôi đặt α thành 5. Sau đó, chúng tôi sử dụng Gumbel-Softmax để có được xác suất µt cho việc chỉ định sổ mã, được chỉ định bởi µt = GumbelSof tmax(zt) R K.

Do đó, xác suất để chọn từ mã thứ k ek để lượng tử hóa x được đưa ra bởi x^N_t

$$\mu t,k = \frac{\exp((\ \alpha \cdot dt,k + vt,k)/\tau\ )}{\sum_{i=1}^{K} biểu\ thức((\ \alpha \cdot dt,i + vt,i)/\tau\ )}, \qquad (5)$$

trong đó τ là nhiệt độ của Gumbel-Softmax, được ủ theo hàm mũ từ 2 đến 0,5 trong thí nghiệm của chúng tôi, và vt,k Gumbel(0, 1) là các mẫu được rút ra từ phân phối Gum-bel. Trong quá trình truyền tới, chỉ số cứng argmaxk {1,2,..,K}µt,k được chọn; do đó, không có khoảng cách giữa quá trình đào tạo và suy luận. Trong quá trình truyền ngược, gradient liên quan đến logit zt được sử dụng.

Ước tính entropy và kiểm soát tốc độ Vì entropy đóng vai trò là giới hạn dưới của tốc độ bit thực tế, chúng tôi tận dụng ước tính entropy để kiểm soát tốc độ bit R(x|0) hướng tới mục tiêu Rtarget đã cho, được thúc đẩy bởi công trình trong [10], [32]. Sử dụng VQ dựa trên Distance-Gumbel-Softmax, chúng tôi có thể tính toán phân phối gán mềm mẫu µ˜, bằng cách cộng các xác suất Softmax cho mỗi từ mã trong một minibatch µ˜t,k,b như

$$\tilde{\mu}k = \frac{1}{Bt}\sum_{b=1}^{B}\sum_{t=1}^{T} \tilde{\mu}t,k,b, \quad k \quad \{1, 2, ..., K\}, \qquad (6)$$

trong đó B và T lần lượt biểu diễn kích thước lô và số khung hình trong mỗi clip âm thanh và µ˜t,k,b

Bài viết này đã được chấp nhận để xuất bản trong IEEE/ACM Transactions on Audio, Speech and Language Processing. Đây là phiên bản của tác giả chưa được biên tập đầy đủ và nội dung có thể thay đổi trước khi xuất bản cuối cùng. Thông tin trích dẫn: DOI 10.1109/TASLP.2023.3277693

là xác suất phân phối Softmax được đưa ra bởi μ˜t,k,b = Sof tmax(zt,k,b) trên K mục nhập codebook. Sau đó, chúng ta có thể ước tính "entropy mềm" trên phân phối gán mềm μ˜ R như

$$H(\tilde{\mu}) \approx \sum_{k=1}^{K} \tilde{\mu}_k \log \tilde{\mu}_k. \quad (7)$$

Kiểm soát tỷ lệ được thực hiện trên mỗi lô nhỏ với hàm mất mát sau Lrate:

$$\text{Tỷ lệ } L = |\text{Mục tiêu } R \quad H(\tilde{\mu})|. \quad (8)$$

Tổn thất Lrate này không chỉ hạn chế tốc độ bit mà còn thực hiện tối ưu hóa tốc độ-biến dạng theo LRD = Dist(x, xˆ) + λ · Lrate. Khi entropy hiện tại cao hơn Rtarget, nó sẽ đẩy các tính năng tương tự được lượng tử hóa thành cùng một từ mã thông qua sự đánh đổi giữa tốc độ và độ biến dạng; trong khi khi nó thấp hơn Rtarget, các tính năng tương tự có thể được lượng tử hóa thành các từ mã khác nhau để duy trì chất lượng cao hơn nhưng ở tốc độ cao hơn. Cần lưu ý rằng mặc dù có một số ước tính ở đây, chúng tôi thấy rằng tốc độ bit thực tế được kiểm soát tốt trong quá trình thử nghiệm.

Để giảm kích thước sổ mã để dễ dàng tối ưu hóa, lượng tử hóa vectơ nhóm [31] được sử dụng. Cụ thể, mỗi khung R Cd được chia thành G nhóm dọc theo kênh di- 'N G ×G và mỗi nhóm được x N t lượng tử hóa, tạo ra x tized với một sổ mã riêng biệt chứa K từ mã {e G ×K. Hơn nữa, bốn t khung chồng chéo Cd R

, 1 , e´ 2 , ..., e' K} R Đĩa CD

(dữ liệu mới 40 ms) được hợp nhất thành một để lượng tử hóa nên kích thước từ mã là Cd × 4 cho mỗi sổ mã. Không giống như các thiết lập sổ mã thông thường G trong các codec thần kinh hiện có [8] [36], chúng tôi thiết lập một kích thước sổ mã lớn hơn để nó có thể nắm bắt được sự phân phối thực sự của các tính năng tiềm ẩn thông qua quá trình tối ưu hóa tốc độ méo. Ví dụ, ở 3 kbps, mỗi dữ liệu mới 40 ms dự kiến sẽ tiêu thụ 120 bit. Các tham số sổ mã G và K được đặt thành 16 và 1024 tương ứng, trong đó G · log2 (K) = 16 · log2 (1024) = 160 > 120. Sau đó, tốc độ bit thực được kiểm soát bởi Công thức 8 để đạt được 3 kbps.

Điều này khá khác so với mất mát đa dạng trong phương pháp dựa trên Gumbel-Softmax [30], trong đó phân phối đồng đều được áp dụng cho việc sử dụng từ mã. Bảng I hiển thị các cấu hình sổ mã G và K ở nhiều tốc độ bit khác nhau trong thí nghiệm của chúng tôi.

F. Đào tạo đối nghịch Đào

tạo đối nghịch đã được chứng minh là rất hiệu quả trong việc tạo ra giọng nói chất lượng cao [37] [38]. Đối với chất lượng nhận thức được tái tạo cao, chúng tôi cũng sử dụng đào tạo đối nghịch trong lược đồ của mình với bộ phân biệt miền tần số. Nó lấy phổ tần số thời gian phức tạp của dạng sóng đầu vào làm đầu vào. Phổ độ lớn được nén theo luật lũy thừa với lũy thừa 0,3 để cân bằng tầm quan trọng tương đối của các thành phần khác nhau. Pha được giữ nguyên. Bốn lớp tích chập 2D với kích thước hạt nhân là (2, 3) và bước nhảy là (2, 2) được sử dụng để trích xuất các tính năng có độ phân giải giảm dần theo cả chiều tần số và thời gian.

Số kênh lần lượt là 8, 8, 16 và 16. Mỗi lớp tích chập được theo sau bởi một chuẩn hóa thể hiện (IN) và một Leaky ReLu [39]. Một phép biến đổi tuyến tính được sử dụng để gộp tất cả thông tin tần số vào các kênh và giảm kích thước kênh xuống còn 1. Cuối cùng, chúng tôi sử dụng một lớp gộp trung bình thời gian với kích thước hạt nhân là 10 và tạo ra các logit một chiều lấy mẫu xuống cuối cùng có kích thước Td trong chiều thời gian.

Chúng tôi sử dụng tổn thất bình phương nhỏ nhất làm mục tiêu đối nghịch, tương tự như trong LSGAN [40]. Tổn thất đối nghịch cho máy phát điện G là

$$\text{Ladv} = \text{Ex}\hat{}[\frac{1}{\text{Td}}\sum_{t=1}^{\text{Td}}(\text{Dt}(\hat{}x) \quad 1)2\ ]. \quad (9)$$

trong đó xˆ = G(x) là tín hiệu được tái tạo. Sự mất mát cho bộ phân biệt D là

$$\text{LD} = \text{Ví dụ}[\frac{1}{\text{Td}}\sum_{t=1}^{\text{Td}}(\text{Dt}(x) \quad 1)2\ ] + \text{Ex}\hat{}[\frac{1}{\text{Td}}\sum_{t=1}^{\text{Td}}(\text{Dt}(x))2\ ]. \quad (10)$$

Chúng tôi cũng sử dụng một đặc điểm mất mát Lf eat để hướng dẫn đào tạo máy phát điện cho chất lượng nhận thức cao [37] [38]. Nó được tính là sự khác biệt ℓ1 của các đặc điểm sâu từ bộ phân biệt giữa âm thanh được tạo ra và âm thanh gốc, được đưa ra bởi

$$\text{Nếu ăn} = \text{Ví dụ}[\frac{1}{\text{L}}\sum_{l=1}^{\text{L}}\frac{1}{\text{ClFlTl}}||\text{Dl}(x) \quad \text{Dl}(\hat{}x)||1], \quad (11)$$

trong đó Dl ,l {1, 2, ..., L} là bản đồ đặc trưng của lớp thứ l của bộ phân biệt, và Cl là độ phân , loa hồng ,Tl biểu thị kênh, tần số giải thời gian và độ phân giải tính năng của Dl . Chúng tôi tính toán độ mất tính năng trên bốn lớp tích chập 2D đầu tiên của bộ phân biệt.

G. Hàm mục tiêu Chúng tôi sử

dụng hàm mất mát sau để hướng dẫn quá trình đào tạo nhằm đạt được chất lượng âm thanh đầu ra tối đa ở tốc độ bit mục tiêu. Tổng mất mát cho trình tạo bao gồm một thuật ngữ tái tạo Lrecon, một thuật ngữ ràng buộc tốc độ Lrate, một thuật ngữ dự đoán Lpred, một thuật ngữ đối nghịch Ladv và một thuật ngữ khớp tính năng Lf eat, tức là

$$\text{LG} = \text{Lrecon} + \lambda\text{rateLrate} + \lambda\text{predLpred} + \lambda\text{advLadv} + \lambda\text{f eatLf eat}, \quad (12)$$

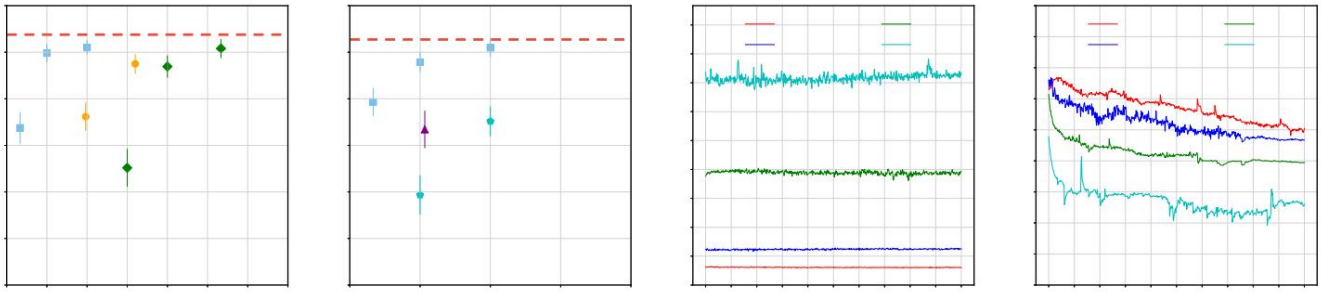
trong đó Lrate, Lpred, Ladv và Lf eat đã được giải thích trong Eq. 8, 3, 9 và 11, tương ứng. Thuật ngữ tái tạo được chọn để đạt được cả độ trung thực tín hiệu cao và chất lượng nhận thức cao. Chúng tôi sử dụng hai thuật ngữ miền tần số cho Lrecon, như được hiển thị bên dưới

$$\text{Lrecon} = \text{Lbin} + \lambda\text{melLmel}. \quad (13)$$

Thuật ngữ đầu tiên Lbin là tổn thất lỗi bình phương trung bình (MSE) trên phổ STFT nén theo luật lũy thừa [41]. Để duy trì tính nhất quán của STFT [42], phổ được tái tạo trước tiên được chuyển đổi sang miền thời gian và sau đó sang miền tần số để tính toán tổn thất. Thuật ngữ thứ hai Lmel là tổn thất phổ mel đa thang đo được đưa ra bởi

$$\text{Lmel} = \text{Ví dụ}[\frac{1}{\text{T}}\sum_{n=1}^{\text{T}}\frac{1}{\text{TnS}}||\varphi\ \text{n}(x) \quad \varphi\ \text{n}(\hat{}x)||1], \quad (14)$$

Bài viết này đã được chấp nhận để xuất bản trong IEEE/ACM Transactions on Audio, Speech and Language Processing. Đây là phiên bản của tác giả chưa được biên tập đầy đủ và nội dung có thể thay đổi trước khi xuất bản cuối cùng. Thông tin trích dẫn: DOI 10.1109/TASLP.2023.3277693



(a) TF-Codec so với codec chuẩn (b) Bộ giải mã TF so với bộ giải mã thần kinh (c) Entropy (d) Sức mạnh có thể học được

Hình 6. (a)(b) Kết quả đánh giá chủ quan. Đường chấm đỏ biểu thị điểm của tham chiếu. Thanh lỗi biểu thị khoảng tin cậy 95%. Chúng tôi sử dụng Opus-1.3.1 và EVS-16.2.0 từ bản phát hành chính thức và đặt chúng ở chế độ WB. Âm thanh được lấy mẫu ở 16 kHz cho các codec này, ngoại trừ Encodec hoạt động ở 24 kHz. Đối với Encodec, âm thanh 16khz được lấy mẫu lên 24 kHz. Bộ dự đoán dựa trên tích chập được sử dụng cho TF-Codec trong thử nghiệm này. Cần lưu ý rằng TF-Codec và Opus hoạt động theo cách bitrate thay đổi (VBR). Cả Lyra-v2 và Encodec đều hoạt động ở chế độ bitrate không đổi (CBR) trong hình này.
EVS được đặt ở chế độ bitrate biến đổi được kiểm soát nguồn (SC-VBR) [44] ở 5,9 kbps và CBR ở 9,6 kbps. (c) Entropy cho dữ liệu 40ms trong quá trình đào tạo. (d) Hệ số công suất đã học trong quá trình đào tạo. Trong (c) và (d), các đường cong chỉ tương ứng với giai đoạn đào tạo đối nghịch.

trong đó φ n(·) là hàm biến đổi dạng sóng thành phổ đồ mel bằng cách sử dụng kích thước cửa sổ thứ n Tn. S biểu thị số lượng thùng mel, được đặt thành 64 cho tất cả các độ dài cửa sổ.
Tiếp theo [43], chúng tôi tính toán phổ mel trên một chuỗi các chiều dài cửa sổ từ 64 đến 2048. Chúng tôi đặt λrate = 0,04, λpred = 0,02, λadv = 0,001, λf eat = 0,1 và λmel = 0,25 để cân bằng các điều khoản khác nhau trong các thí nghiệm của chúng tôi.

IV. KẾT QUẢ THỰC NGHIỆM

Trong phần này, chúng tôi đánh giá TF-Codec được đề xuất dựa trên công nghệ tiên tiến nhất và cung cấp phân tích chi tiết về từng phần để chứng minh những gì nó học được và lý do tại sao nó hoạt động hiệu quả.

A. Bộ dữ liệu và Cài đặt

Chúng tôi lấy 890 giờ giọng nói trong trẻo 16kHz từ Thử thách giảm tiếng ồn sâu tại ICASSP 2021 [45], bao gồm các đoạn clip giọng nói, cảm xúc và ca hát đa ngôn ngữ.
Mỗi âm thanh được cắt thành các clip 3 giây với mức độ giọng nói ngẫu nhiên từ [50, 10] db để đào tạo. Để đánh giá, chúng tôi sử dụng 1458 clip 10 giây mà không có bất kỳ sự chồng chéo nào với dữ liệu đào tạo, bao gồm hơn 1000 người nói với nhiều ngôn ngữ. Một cửa sổ Hanning được sử dụng trong STFT với độ dài cửa sổ là 40 ms và độ dài bước nhảy là 10 ms.

Tất cả các mô-đun của TF-Codec, bao gồm mã hóa, giải mã và lượng tử hóa, có thể được đào tạo từ đầu đến cuối trong một giai đoạn duy nhất. Đối với đào tạo đối nghịch, trước tiên chúng tôi đào tạo một máy phát điện tốt từ đầu đến cuối và sau đó tinh chỉnh máy phát điện bằng một bộ phân biệt theo cách đối nghịch.

Trong quá trình đào tạo, chúng tôi sử dụng trình tối ưu hóa Adam [46] với tốc độ học là 3 × 10 4 cho trình tạo ở giai đoạn đầu tiên. Sau đó, trình tạo và bộ phân biệt được đào tạo với tốc độ học lần lượt là 3 × 10 5 và 3 × 10 4 . Chúng tôi đào tạo cả hai giai đoạn trong 100 kỷ nguyên với kích thước lô là 100.

B. So sánh với các bộ giải mã hiện đại

Đầu tiên, chúng tôi so sánh TF-Codec được đề xuất với một số codec truyền thống và hai codec thần kinh mới nhất để chứng minh khả năng biểu diễn mạnh mẽ của xương sống của chúng tôi. Chúng tôi

BẢNG I CẤU
HÌNH SỔ MÃ VÀ PHÂN TÍCH TỐC ĐỘ BITRÊN BỘ KIỂM TRA.

Chế độ Bitrate (kbps)	Kích thước sổ mã Huffman mã hóa G · log2 (K)	Tốc độ bit trung bình (kbps)
0,512	G=3, K=512	0,498
	G=6, K=1024	1,014
	G=16, K=1024	3,089
1 3 6	G=32, K=1024	6,162

tiến hành một bài kiểm tra nghe chủ quan bằng phương pháp lấy ý kiến cộng đồng lấy cảm hứng từ MUSHRA [47], trong đó 10 người tham gia đánh giá 15 mẫu từ bộ kiểm tra. Trong quá trình đánh giá MUSHRA, người nghe được cung cấp một tham chiếu ẩn và một bộ mẫu kiểm tra được tạo ra bởi các phương pháp khác nhau. Mỏ neo được lọc thông thấp không được sử dụng trong thí nghiệm của chúng tôi.

Chúng tôi lấy hai codec chuẩn, tức là Opus và EVS để so sánh. Opus1 [28] là một codec đa năng được sử dụng rộng rãi cho truyền thông thời gian thực, hỗ trợ giọng nói và âm thanh băng tần hẹp đến băng tần đầy đủ với tốc độ bit từ 6 kbps đến 510 kbps.
Bộ giải mã EVS [44] được 3GPP phát triển và chuẩn hóa chủ yếu cho Thoại qua LTE (VoLTE). Chúng tôi cũng so sánh với hai bộ giải mã thần kinh mới nhất, tức là Lyra-v2 và Encodec [36] làm việc đồng thời. Lyra-v22 là phiên bản cải tiến của Lyra-v1, tích hợp kiến trúc của SoundStream [8] để có chất lượng âm thanh tốt hơn và khả năng mở rộng tốc độ bit3 . Encodec4 tạo ra âm thanh có độ trung thực cao trên nhiều loại băng thông và loại âm thanh.

Hình 6 (a)(b) cho thấy kết quả đánh giá, trong đó chúng tôi so sánh TF-Codec của mình từ 1 kbps đến 6 kbps với các đường cơ sở codec chuẩn và thần kinh ở nhiều tốc độ bit khác nhau. Người ta quan sát thấy

rằng TF-Codec của chúng tôi ở tốc độ 1 kbps vượt trội hơn đáng kể so với Opus ở tốc độ 9 kbps, Lyra-v2 ở tốc độ 3,2 kbps và EnCodec ở tốc độ 6 kbps, chứng tỏ khả năng biểu diễn mạnh mẽ của TF-

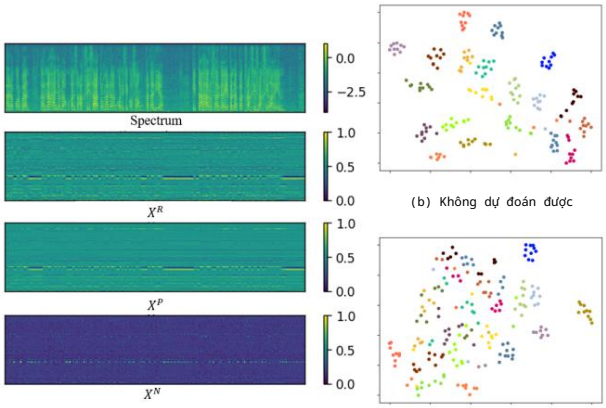
1https://opus-codec.org
2https://github.com/google/lyra
3https://opensource.googleblog.com/2022/09/lyra-v2-a-better-faster-and-more-versatile-speech-codec.html
4https://github.com/facebookresearch/encodec

Bài viết này đã được chấp nhận để xuất bản trong IEEE/ACM Transactions on Audio, Speech and Language Processing. Đây là phiên bản của tác giả chưa được biên tập đầy đủ và nội dung có thể thay đổi trước khi xuất bản cuối cùng. Thông tin trích dẫn: DOI 10.1109/TASLP.2023.3277693

BẢNG II ĐÁNH

GIÁ VỀ MÃ HÓA DỰ ĐOÁN Ở 3 KBPS(W/O. ADV).

Phương pháp	PESQ	STOI	ViSQOL
Bộ giải mã TF không có Dự đoán	2,763	0,917	Bộ giải mã TF 3.219
có Thích ứng	2,774	0,914	Bộ giải mã TF có Chuyển đổi 3.332
	2,895	0,917	3.345



(a) Hình ảnh tính năng

(b) Không dự đoán được

(c) Dự đoán

Hình 7. Hình ảnh trực quan về tính năng của mã hóa dự đoán. (a) Bốn hàng từ trên xuống dưới hiển thị phổ STFT (log-scale) của âm thanh chưa nén, đầu ra của bộ mã hóa trước khi mã hóa dự đoán, đầu ra của bộ dự đoán và đầu ra của bộ trích xuất. Các giá trị được chuẩn hóa tuyến tính giữa 0 và 1. (b)(c) Hình ảnh trực quan T-SNE của thông tin người nói bằng mã hóa không dự đoán và dự đoán ở tốc độ 3 kbps.

Codec. Khi hoạt động ở tốc độ 3 kbps, TF-Codec của chúng tôi đạt hiệu suất tốt hơn EVS ở mức 9,6 kbps và Opus ở mức 12 kbps, đồng thời cũng vượt trội hơn hẳn hai bộ giải mã thần kinh. Ở dải bitrate cao hơn, TF-Codec của chúng tôi ở tốc độ 6 kbps có hiệu suất ngang bằng với Opus ở tốc độ 16 kbps và vượt trội hơn Encodec ở tốc độ 6 kbps ở biên độ lớn. Bên cạnh đó, Hình 6 (c) cho thấy tổng entropy của TF-Codec của chúng tôi được kiểm soát với tốc độ mất Lrate trong quá trình đào tạo. Bảng I hiển thị tốc độ bit thực tế sau khi mã hóa Huffman trên bộ thử nghiệm. Chúng ta có thể thấy rằng tốc độ bit thực tế được kiểm soát tốt. Vì mã hóa độ dài thay đổi không thể đảm bảo bit không đổi cho mỗi khung hình, nên chúng tôi cũng thu thập số liệu thống kê về mức tiêu thụ bit trên mỗi khung hình. Ở tốc độ 3kbps, bit tối đa, tối thiểu và trung bình trên tất cả các khung hình của toàn bộ bộ thử nghiệm lần lượt là 190, 39 và 124. Đối với một âm thanh duy nhất, sự thay đổi nhỏ hơn với độ lệch chuẩn trung bình là 40 và tối đa là 53.

C. Nghiên cứu cắt bỏ

Để đánh giá các phần khác nhau của phương pháp được đề xuất, chúng tôi sử dụng một số số liệu khách quan, bao gồm PESQ băng rộng [48], STOI [49] và VISQOL [50]. Mặc dù các số liệu này không được thiết kế và tối ưu hóa cho cùng một nhiệm vụ, chúng tôi thấy rằng đối với cùng một loại biến dạng trong tất cả các lược đồ được so sánh, chúng phù hợp với chất lượng nhận thức.

1) Mã hóa dự đoán: Đầu tiên, chúng tôi đánh giá hiệu quả của mã hóa dự đoán. Chúng tôi so sánh hai biến thể của TF-Codec theo các bộ dự đoán dựa trên tích chập và thích ứng, tương ứng, với biến thể không có mã hóa dự đoán bằng cách vô hiệu hóa vòng lặp dự đoán. Bảng II hiển thị kết quả đánh giá trong đó việc đào tạo đối nghịch bị vô hiệu hóa đối với tất cả các phương pháp so sánh. Có thể thấy rằng khi tất cả hoạt động ở tốc độ 3kbps, mã hóa dự đoán cải thiện chất lượng âm thanh được tái tạo trong cả PESQ và ViSQOL với độ rõ lời nói tương tự như được đo bằng STOI. Phương pháp dựa trên tích chập vượt trội hơn cơ chế thích ứng vì sau khi lượng tử hóa, giả định về dự đoán tuyến tính hằng số cục bộ có thể không còn đúng nữa trong lược đồ thích ứng.

Để tìm hiểu sâu hơn về các biểu diễn mà nó học được, chúng tôi trực quan hóa các tính năng của các mô-đun khác nhau bằng mã hóa dự đoán trong Hình 7 (a). Bốn hàng từ trên xuống dưới hiển thị phổ STFT của âm thanh chưa nén, đầu ra của bộ mã hóa XR trước khi mã hóa dự đoán, đầu ra của bộ dự đoán XP và đầu ra của bộ trích xuất XN. Có thể thấy rằng XP dự đoán khá giống với XR, cho thấy bộ dự đoán cung cấp dự đoán tốt về khung hiện tại từ quá khứ. Chúng tôi cũng có thể quan sát thấy rằng tính năng XN sau khi bộ trích xuất trở nên thưa thớt hơn nhiều so với XR, cho thấy rằng hầu hết thông tin dư thừa đã bị loại bỏ. Chúng tôi cũng tính toán hệ số tương quan thời gian của biểu diễn đã học, tức là đầu ra lớp cuối cùng của mã hóa. Mã hóa không dự đoán mà không có vòng lặp dự đoán đạt được hệ số tương quan trung bình là 0,37, trong khi mã hóa dự đoán giảm hệ số này xuống 0,09, cho thấy rằng tương quan thời gian bị loại bỏ triệt để hơn trong mã hóa dự đoán. Điều này cũng phù hợp với kết quả trực quan hóa trong Hình 7 (a).

Để khám phá thêm thông tin dư thừa nào đã bị loại bỏ bằng mã hóa dự đoán, chúng tôi trình bày trực quan hóa t-SNE [51] của thông tin người nói có trong các biểu diễn đã học trong Hình 7(b)(c) cho mã hóa không dự đoán và mã hóa dự đoán, tương ứng. Để đạt được điều này, chúng tôi thực hiện nhóm thời gian trên biểu diễn đã học, tạo ra một vectơ nhúng cho mỗi âm thanh. Các phát ngôn của 20 người nói được chọn ngẫu nhiên từ tập dữ liệu Librispeech [52] được sử dụng để trực quan hóa. Chúng tôi có thể quan sát thấy rằng các biểu diễn từ mã hóa không dự đoán được nhóm lại tốt cho mỗi người nói, cho thấy rằng chúng chứa hầu hết thông tin của người nói. Ngược lại, trong mã hóa dự đoán, các nhúng phân tán đối với hầu hết người nói, cho thấy thông tin của người nói đã bị loại bỏ hiệu quả.

Điều này hợp lý vì thông tin liên quan đến người nói tương đối ổn định theo thời gian và dễ dự đoán.

2) Nén đầu vào có thể học được: Để chứng minh hiệu quả của nén đầu vào có thể học được, chúng tôi so sánh nó với nén theo luật lũy thừa cố định trong đó tham số công suất được đặt thành 0,3 như trong [41]. Bảng III cho thấy ở 1 kbps, nén có thể học được rõ ràng vượt trội hơn nén cố định ở cả ba số liệu, thể hiện cả chất lượng nhận thức tốt hơn và khả năng hiểu giọng nói tốt hơn. Trong đánh giá chủ quan của mình, chúng tôi cũng thấy chất lượng nhận thức tăng rõ ràng đối với tốc độ bit rất thấp, chẳng hạn như 1 kbps. Để tìm hiểu các tham số công suất mà nó học được, chúng tôi cũng báo cáo công suất đã học được p ở nhiều tốc độ bit khác nhau trong quá trình đào tạo, như thể hiện trong Hình 6 (d). Chúng ta có thể thấy rằng công suất đã học được giảm dần trong quá trình đào tạo, cho thấy rằng mô hình đầu tiên chủ yếu tập trung vào các thùng năng lượng cao, thường là các băng tần số thấp. Khi kỷ nguyên tăng lên, mô hình chuyển sang chú ý nhiều hơn đến các chi tiết năng lượng thấp của quang phổ. Người ta cũng quan sát thấy rằng tốc độ bit càng cao thì p càng nhỏ, điều đó có nghĩa là mô hình cố gắng kiểm tra

Bài viết này đã được chấp nhận để xuất bản trong IEEE/ACM Transactions on Audio, Speech and Language Processing. Đây là phiên bản của tác giả chưa được biên tập đầy đủ và nội dung có thể thay đổi trước khi xuất bản cuối cùng. Thông tin trích dẫn: DOI 10.1109/TASLP.2023.3277693

BẢNG III
ĐÁNH GIÁ VỀ NÉN ĐẦU VÀO CÓ THỂ HỌC ĐƯỢC Ở TỐC ĐỘ 1 KBPS.

Phương pháp	PESQ	STOI	ViSQOL
nén cố định	2.289	0.877	2.781
nén có thể học được	2.351	0.887	2.851

BẢNG IV
ĐÁNH GIÁ VỀ LƯỢNG TỬ HÓA VECTOR Ở TỐC ĐỘ 3 KBPS.

Phương pháp	PESQ	STOI	ViSQOL
Gumbel-Softmax	2.738	0.910	3.204
Khoảng cách-Gumbel-Softmax	2.763	0.917	3.219

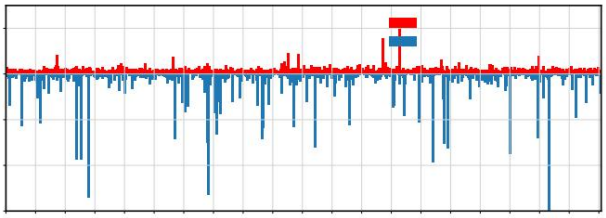
các thành phần chi tiết với nhiều bit hơn có sẵn, tạo ra chất lượng nhận thức tốt hơn.

3) VQ dựa trên Distance-Gumbel-Softmax: Chúng tôi so sánh Cơ chế VQ dựa trên Distance-Gumbel-Softmax với phương pháp dựa trên Gumbel-Softmax trước đó trong [30]. Bảng IV cho thấy ở tốc độ 3 kbps, phương pháp của chúng tôi vượt trội hơn phương pháp dựa trên Gumbel-Softmax trước đây ở mọi số liệu, cho thấy việc đưa thông tin khoảng cách vào ví dụ giúp cải thiện chất lượng tái tạo.

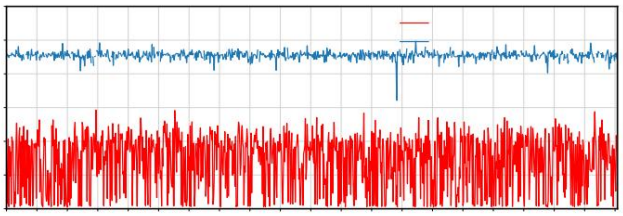
Chúng tôi cũng cho thấy sự phân bố của các sổ mã đã học để giúp hiểu cách thức Distance-Gumbel-Softmax dựa trên lượng tử hóa vector học. Hình 8(a) cho thấy cách sử dụng 1024 từ mã của một sổ mã trên bộ thử nghiệm với 1458 âm thanh đối với cả phương pháp dựa trên Gumbel-Softmax [30] và cơ chế Distance-Gumbel-Softmax được đề xuất. Chúng ta có thể quan sát rằng các từ mã có xu hướng được phân bố đồng đều hơn trong phương pháp dựa trên Gumbel-Softmax, trong khi ở phương pháp Distance-Gumbel-Softmax, các từ mã được phân phối đa dạng hơn, với một số từ mã được sử dụng rất thường xuyên.
Điều này là hợp lý, như trong phương pháp dựa trên Gumbel-Softmax, mất đa dạng Sự đa dạng được áp đặt lên các từ mã đã học, điều này khuyến khích mỗi từ mã được sử dụng bình đẳng. Trong ngược lại, trong Distance-Gumbel-Softmax, chúng tôi sử dụng một codebook và sử dụng tốc độ mất mát Lrate để đạt được tốc độ bit mục tiêu theo nghĩa tối ưu hóa tỷ lệ biến dạng. Theo cách này, thực tế sự phân bố các đặc điểm tiềm ẩn có thể được nắm bắt trong Distance-Gumbel-Softmax.

Chúng tôi cũng hiển thị điểm tin cậy của việc lựa chọn tốt nhất từ mã trong Hình 8(b), dựa trên xác suất mềm đã học bởi softmax. Khoảng cách-Gumbel-Softmax hiển thị rõ ràng độ tin cậy cao hơn nhiều so với Gumbel-Softmax, điều này chỉ ra rằng sổ mã đã học trong Distance-Gumbel-Softmax có nhiều trung tâm lớp học riêng biệt hơn. Điều này là do giới thiệu bản đồ khoảng cách, tức là lỗi lượng tử hóa, vào xác suất mềm, theo đó từ mã gần hơn với tính năng hiện tại được khuyến khích để được lựa chọn, và rõ ràng hơn các từ mã được học thông qua quá trình lan truyền ngược.

4) Huấn luyện đối nghịch: Chúng tôi cũng tiến hành một nghiên cứu cắt bỏ để đánh giá hiệu suất của đào tạo đối kháng, như được trình bày trong Bảng V. Chúng tôi trình bày sự so sánh có và không có đào tạo đối nghịch ở nhiều tốc độ bit khác nhau. Người ta quan sát thấy rằng với đào tạo đối kháng, PESQ, STOI và ViSQOL phần lớn là được cải thiện, đặc biệt là ở tốc độ 1 kbps và 3 kbps. Chúng tôi cũng quan sát rằng sự mất mát phù hợp với đặc điểm Lf eat đóng vai trò quan trọng



(a) Biểu đồ sử dụng từ mã



(b) Độ tin cậy của từ mã

Hình 8. Đặc điểm của một sổ mã học được chọn ngẫu nhiên ở tốc độ 3 kbps. (a) Tần suất 1024 từ mã được chọn. (b) Điểm tin cậy của 1024 từ mã.

BẢNG V
ĐÁNH GIÁ VỀ ĐÀO TẠO ĐỐI KHÁNG.

Phương pháp	Tốc độ bit	PESQ	STOI	ViSQOL
Bộ giải mã TF không có Adv.	1 kbps	2.085	0.868	2.742
Bộ giải mã TF có Adv.	1 kbps	2.351	0.887	2.851
Bộ giải mã TF không có Adv.	3 kbps	2,763	0,917	3.219
Bộ giải mã TF có Adv.	3 kbps	3,124	0,933	3.510
Bộ giải mã TF không có Adv.	6 kbps	3,426	0,949	3.966
Bộ giải mã TF có Adv.	6 kbps	3,547	0,953	3.841

trong việc khôi phục các chi tiết tần số cao trong âm thanh được tạo ra trong thí nghiệm của chúng tôi.

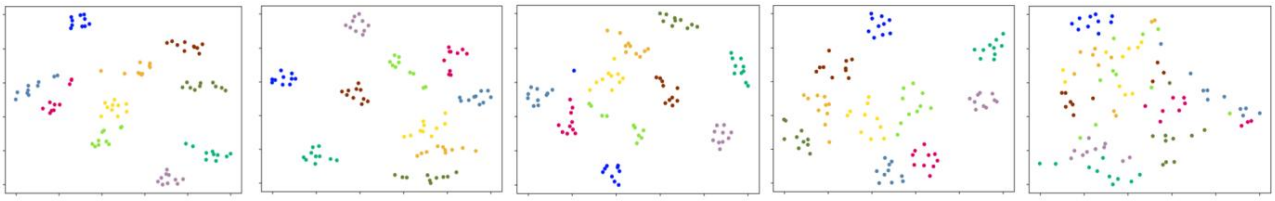
D. Phân tích

Để hiểu rõ hơn thông tin nào được học và mã hóa ở nhiều tốc độ bit khác nhau, chúng tôi tiến hành phân tích thông tin đã học biểu diễn và sổ mã trong phần này. Trong phân tích này, chúng tôi vô hiệu hóa vòng lặp dự đoán và chọn tiềm ẩn rời rạc mã để trực quan hóa.

Chúng tôi hình dung người nói và thông tin nội dung chứa đựng trong các mã tiềm ẩn rời rạc đã học được bằng cách sử dụng hai tập dữ liệu: (i) bộ dữ liệu nhiều người nói Librispeech [52] cho người nói liên quan phân tích thông tin; (ii) tập dữ liệu người nói đơn LJ [54] cho phân tích thông tin ngôn ngữ.

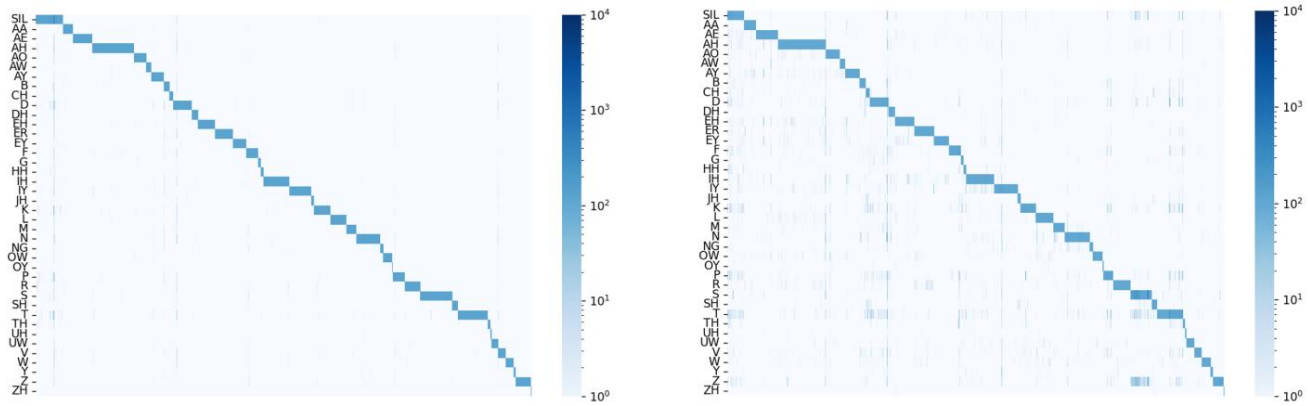
Thông tin diễn giả Chúng tôi chọn ngẫu nhiên 10 diễn giả từ Librispeech, mỗi bài có 10 câu nói. Đối với mỗi câu nói, chúng tôi thực hiện một nhóm trung bình thời gian trên các tính năng đa khung, tạo ra một nhúng toàn cầu cho mỗi phát ngôn. Hình 9 cho thấy hình ảnh t-SNE của các nhúng loa đó từ 0,256 kbps đến 6 kbps. Chúng ta có thể quan sát thấy mô hình ở mức cao tốc độ bit tạo ra các cụm loa nhỏ gọn hơn, trong khi ở mức rất tốc độ bit thấp, cụm bắt đầu khuếch tán và loa có thể

Bài viết này đã được chấp nhận để xuất bản trong IEEE/ACM Transactions on Audio, Speech and Language Processing. Đây là phiên bản của tác giả chưa được biên tập đầy đủ và nội dung có thể thay đổi trước khi xuất bản cuối cùng. Thông tin trích dẫn: DOI 10.1109/TASLP.2023.3277693



(a) 6 kbps (b) 3kbps (c) 1kbps (d) 0,512 kbps (e) 0,256 kbps

Hình 9. T-SNE của thông tin người nói trong các mã tiềm ẩn rời rạc trên tập dữ liệu Librispeech.



(a) 3kbps (b) 0,512 kbps

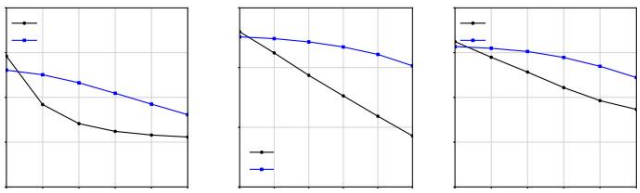
Hình 10. Sự đồng hiện của mã tiềm ẩn và âm vị trên tập dữ liệu người nói đơn LJ. Trục ngang là chỉ số mã tiềm ẩn và trục dọc biểu thị âm vị. Chúng tôi thu được các căn chỉnh cấp âm vị với Montreal Forced Aligner (MFA) [53], sử dụng mô hình âm thanh Librispeech được đào tạo trước của họ. Các nhãn âm vị cấp khung được xác định bởi âm vị có nhiều lần xuất hiện nhất trong thời lượng của mỗi khung.

không được xác định ở 0,256 kbps. Điều này chỉ ra rằng ở tốc độ bit rất thấp, mô hình chuyển sang loại bỏ thông tin liên quan đến người nói để lại băng thông cho một số thông tin chính của giọng nói. Điều đáng chú ý là tốc độ bit 0,256 kbps thậm chí còn gần với tốc độ thông tin ước tính của giao tiếp giọng nói trong [55].
Ở tốc độ bit thấp như vậy, thông tin ngôn ngữ quan trọng hơn sự khác biệt về giọng nói trong giao tiếp thời gian thực.

Thông tin ngôn ngữ Chúng tôi đánh giá thông tin ngôn ngữ trong các mã rời rạc bằng bản đồ đồng hiện giữa các âm vị và các mã tiềm ẩn rời rạc bằng lượng tử hóa vectơ nhóm. Chúng tôi sử dụng tất cả 13100 đoạn âm thanh của tập dữ liệu LJSpeech do cùng một người nói để loại bỏ tác động do sự thay đổi của người nói. Hình 10 cho thấy bản đồ đồng hiện ở các tốc độ bit khác nhau. Có thể thấy rằng các mã tiềm ẩn rời rạc này, được học theo cách tự giám sát, có liên quan chặt chẽ đến các âm vị và nhiều mã tiềm ẩn được dành riêng cho các âm vị cụ thể. Ví dụ, một số lượng lớn các từ mã rời rạc được tự động phân bổ cho các âm vị cụ thể, ví dụ: AH, N, S và T. Người ta cũng quan sát thấy rằng các phân phối của bản đồ đồng hiện cho 0,512 kbps và 3 kbps khá gần nhau, cho thấy rằng các mã tiềm ẩn ở các tốc độ bit khác nhau bảo toàn thông tin âm vị tốt. Điều này phù hợp với giả thuyết của chúng tôi rằng ở tốc độ bit cực thấp, mô hình sẽ cố gắng phân bổ băng thông hạn chế cho thông tin nội dung chính (liên quan đến ngôn ngữ) trong lời nói và loại bỏ thông tin ít quan trọng hơn (liên quan đến người nói).

E. Độ bền vững đối với lỗi truyền dẫn

Người ta thường cho rằng vòng lặp dự đoán được sử dụng để giảm sự dư thừa về mặt thời gian nhạy cảm với lỗi truyền tải,



(a) PESQ (b) STOI (c) PLCMOS

Hình 11. Đánh giá TF-Codec ở tỷ lệ mất gói tin từ 0% đến 50% trên tổng hợp bộ kiểm tra.

vì nó có thể dẫn đến sự lan truyền lỗi dài hạn. Trong phần này, chúng tôi đề xuất một số cách để cải thiện tính mạnh mẽ trong trường hợp mất gói tin với đào tạo nhận biết mất mát và cho thấy một số kết quả sơ bộ nhưng đầy hứa hẹn.

Chúng tôi mô phỏng tình trạng mất gói tin với tỷ lệ mất ngẫu nhiên từ {10%, 20%, 30%, 40%, 50%}, 100 giờ cho mỗi danh mục.
Chúng tôi cũng mô phỏng 390 giờ dữ liệu với mẫu mất gói WLAN bằng cách sử dụng mô hình Markov ba trạng thái, tương tự như trong [56] để đào tạo. Trong trường hợp mất gói, vòng lặp dự đoán ở phía bộ mã hóa hoạt động giống như trước, nhưng trong quá trình giải mã, tính năng lượng tử $\hat{x}$ được đặt thành $\mathbf{x}_{\text{N}}$ không nếu gói bị mất.
Chúng tôi chia mỗi dữ liệu 40ms thành hai gói dọc theo chiều kênh để có khả năng phục hồi tốt hơn. Đối với số liệu đánh giá, bên cạnh PESQ và STOI, chúng tôi cũng sử dụng PLCMOS [57], một công cụ đánh giá được đề xuất trong Thử thách che giấu mất gói tại INTERSPEECH 2022 để đo chất lượng che giấu.

Ngoài việc đào tạo trên các tập dữ liệu mất gói mô phỏng, chúng tôi

Bài viết này đã được chấp nhận để xuất bản trong IEEE/ACM Transactions on Audio, Speech and Language Processing. Đây là phiên bản của tác giả chưa được biên tập đầy đủ và nội dung có thể thay đổi trước khi xuất bản cuối cùng. Thông tin trích dẫn: DOI 10.1109/TASLP.2023.3277693

cũng giới thiệu một thuật ngữ mất mát nhận biết lỗi Lerror aware trên bộ dự đoán trong vòng giải mã, được đưa ra bởi

$$\text{Lerror aware} = \mathrm{Ex}\Big[\frac{1}{CdT}\sum_{t=1}^{T} \|\tilde{x}_t^{P}, \mathrm{sg}(x_t^{R})\|_{2}^{2}\Big], \quad (15)$$

khá giống với Lpred trong Eq. (3) nhưng khác ở chỗ dự đoán $\tilde{x}_t^{P}$ được thực hiện trên tính năng được tái tạo ở phía bộ giải mã với lỗi truyền. Sự mất mát này sẽ buộc bộ dự đoán phải cung cấp một dự đoán có khả năng chống lỗi khi giải mã. Chúng tôi cũng thấy trong thí nghiệm của mình rằng tính năng giống như phần dư để lượng tử hóa x_t^{N} trong TF-Codec dự đoán thường có năng lượng thấp và việc mất các tính năng có năng lượng cực thấp thường không có tác động nghiêm trọng đến quá trình tái tạo.

Hình 11 cho thấy kết quả. Có thể thấy rằng nếu không có quá trình đào tạo nhận biết mất mát, được biểu thị là "Bộ giải mã TF không lỗi" trong Hình 11, chất lượng sẽ giảm mạnh khi có mất gói tin, cho thấy độ nhạy của nó với lỗi truyền. Khi đào tạo với mất gói tin mô phỏng, được biểu thị là "Bộ giải mã TF có khả năng phục hồi lỗi", độ mạnh mẽ và khả năng khôi phục của Bộ giải mã TF được cải thiện đáng kể. Tuy nhiên, âm thanh được tái tạo vẫn có các hiện tượng nhiễu có thể nhận thấy, đặc biệt là đối với mất mát bùng nổ, tức là trên 120ms, có thể được tối ưu hóa thêm.

Điều đáng nói là có nhiều kỹ thuật có thể làm giảm tác động của lỗi truyền trong mã hóa video và âm thanh truyền thống, chẳng hạn như làm mới nội bộ trong H.26X và bảo vệ lỗi chuyển tiếp với các gói dữ liệu dự phòng trong Opus. Trong tương lai, chúng ta có thể xem xét chúng để tối ưu hóa hơn nữa theo hướng mã hóa giọng nói thần kinh có khả năng phục hồi lỗi.

F. Sự chậm trễ và phức tạp

Bộ giải mã TF dự đoán được đề xuất có tổng cộng 6,37M tham số, với 2,11M cho bộ mã hóa 2D, 3,44M cho bộ giải mã 2D và 0,82M cho vòng lặp dự đoán (bao gồm bộ dự đoán, bộ trích xuất và bộ tổng hợp). Chúng tôi báo cáo độ trễ thuật toán và hệ số thời gian thực (RTF) như sau.

Độ trễ của thuật toán Độ trễ của thuật toán đề cập đến độ trễ do phải dựa vào các mẫu trong tương lai để xử lý mẫu hiện tại.
Tất cả các lớp mã hóa và giải mã trong codec của chúng tôi đều mang tính nhân quả, do đó độ trễ của thuật toán xuất phát từ cửa sổ phân tích STFT có độ dài 40 ms, cộng với độ trễ thêm 30 ms khi chúng tôi lượng tử hóa và mã hóa bốn khung hình cùng nhau.

Hệ số thời gian thực RTF được tính là tỷ lệ giữa thời lượng âm thanh và thời gian suy luận. RTF lớn hơn 1 có nghĩa là hệ thống có thể xử lý dữ liệu theo thời gian thực. Khi chạy trên một CPU duy nhất (Intel® Xeon®

Bộ xử lý E5-1620 v4 3,50 GHz), TF-Codec có thể đạt tốc độ mã hóa 4,1 lần và tốc độ giải mã 6,3 lần, đạt được khả năng xử lý thời gian thực.

V. KẾT LUẬN

Chúng tôi đề xuất TF-Codec, một codec giọng nói thần kinh có độ trễ thấp, hoạt động tốt hơn các codec âm thanh tiên tiến với tốc độ bit rất thấp. Chúng tôi đưa mã hóa dự đoán miền tiềm ẩn vào khuôn khổ VQ-VAE để loại bỏ hoàn toàn sự dư thừa về mặt thời gian. Một nén đầu vào có thể học được được đề xuất để cân bằng sự chú ý dành cho các thành phần chính và các chi tiết trong STFT phổ ở các tốc độ bit khác nhau. Chúng tôi cũng giới thiệu cơ chế Distance-Gumbel-Softmax để lượng tử hóa vectơ, có thể nắm bắt được sự phân bố thực tế của các đặc điểm tiềm ẩn với tối ưu hóa tốc độ-biến dạng. Cần lưu ý rằng mặc dù mã hóa giọng nói được lấy làm ví dụ trong bài báo này, các kỹ thuật được đề xuất cũng có thể được mở rộng sang các tín hiệu âm thanh khác như âm nhạc. Trong tương lai, chúng tôi sẽ thực hiện các phần mở rộng như vậy.

Hơn nữa, chúng tôi sẽ nghiên cứu các biểu diễn chi tiết hơn về mặt không chỉ thông tin về người nói và nội dung mà còn về ngữ điệu và cảm xúc. Một điểm thú vị khác cần khám phá trong tương lai là làm cho hệ số nén đầu vào thích ứng với nội dung đầu vào, vì các nội dung khác nhau có tần suất khác nhau

phản hồi.

TÀI LIỆU THAM KHẢO

[1] W. Kleijin, F. Lim, A. Luebs và J. Skoglund, "Mã hóa giọng nói tốc độ thấp dựa trên WaveNet," trong ICASSP. IEEE, 2018, trang 676–680.

[2] WB Kleijn, A. Storus, M. Chinen, T. Denton, FS Lim, A. Luebs, J. Skoglund và H. Yeh, "Mã hóa giọng nói tạo ra với quy tắc hóa phương sai dự đoán," trong ICASSP 2021-2021 IEEE Hội nghị quốc tế về âm học, giọng nói và xử lý tín hiệu (ICASSP). IEEE, 2021, tr. 6478–6482.

[3] J. Klejsa, P. Hedelin, C. Zhou, R. Fejgin và L. Villemoes, "Mã hóa giọng nói chất lượng cao với RNN mẫu," trong ICASSP. IEEE, 2019, trang 7155–7159.

[4] R. Fejgin, J. Klejsa, L. Villemoes và C. Zhou, "Mã hóa nguồn tín hiệu âm thanh bằng mô hình tạo sinh", trong ICASSP. IEEE, 2020, tr. 341–345.

[5] J. Skoglund và J. Valin, "Cải thiện chất lượng tốc độ bit thấp của Opus bằng tổng hợp giọng nói thần kinh," trong Interspeech, 2020.

[6] C. Garbacea, A. van den Oord, Y. Li, F. Lim, A. Luebs, O. Vinyals và TC Walters, "Mã hóa giọng nói tốc độ bit thấp với VQ-VAE và bộ giải mã WaveNet," trong Hội nghị quốc tế IEEE năm 2019 về Xử lý tín hiệu giọng nói âm thanh (ICASSP). IEEE, 2019, trang 735–739.

[7] J. Williams, Y. Zhao, E. Cooper và J. Yamagishi, "Học các biểu diễn điện thoại và loa không đồng nhất trong mô hình VQ-VAE bán giám sát," trong Hội nghị quốc tế IEEE năm 2021 về Xử lý tín hiệu giọng nói âm thanh (ICASSP). IEEE, 2021.

[8] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund và M. Tagliasacchi, "Soundstream: Bộ giải mã âm thanh thần kinh đầu cuối", Giao dịch IEEE/ACM về Xử lý âm thanh, giọng nói và ngôn ngữ, 2021.

[9] K. Zhen, J. Sung, M. Lee, S. Beack và M. Kim, "Học tập dư thừa liên mô-đun theo tầng hướng tới mã hóa giọng nói đầu cuối nhẹ," trong Biên bản Hội nghị thường niên của Hiệp hội giao tiếp và giọng nói quốc tế (Interspeech), 2019.

[10] K. Zhen, J. Sung, MS Lee và M. Kim, "Mã hóa giọng nói thần kinh có thể mở rộng và hiệu quả: một thiết kế kết hợp", Giao dịch IEE/ACM về xử lý âm thanh, giọng nói và ngôn ngữ, tập 30, trang 12–25, 2022.

[11] T. Jayashankar, T. Koehler, K. Kalgaonkar, Z. Xiu, J. Wu, J. Lin, P. Agrawal và Q. He, "Kiến trúc cho bộ giải mã giọng nói thần kinh tốc độ bit thay đổi với độ phức tạp tính toán có thể định cấu hình", trong ICASSP 2022 - Hội nghị quốc tế IEEE về âm học, giọng nói và xử lý tín hiệu (ICASSP), 2022, trang 861–865.

[12] X. Jiang, X. Peng, C. Zheng, H. Xue, Y. Zhang và Y. Lu, "Mã hóa giọng nói thần kinh đầu cuối cho truyền thông thời gian thực," trong IEEE Int. Conf.
Xử lý tín hiệu giọng nói Acoust (ICASSP), 2022.

[13] A. vd Oord, O. Vinyals, và K. Kavukcuoglu, "Rời rạc thần kinh
"học biểu diễn", NIPS, 2017.

[14] GK Wallace, "Tiêu chuẩn nén ảnh tĩnh jpeg," Giao dịch IEEE về điện tử tiêu dùng, tập 38, số 1, trang xviii–xxxiv, 1992.

[15] T. Wiegand, GJ Sullivan, G. Bjontegaard và A. Luthra, "Tổng quan về tiêu chuẩn mã hóa video h. 264/avc," Giao dịch IEEE về mạch và hệ thống cho công nghệ video, tập 13, số 7, trang 560–576, 2003.

[16] GJ Sullivan, J.-R. Ohm, W.-J. Han và T. Wiegand, "Tổng quan về tiêu chuẩn mã hóa video hiệu suất cao (HEVC)," Giao dịch IEEE về mạch và hệ thống cho công nghệ video, tập 22, trang 1649–1668, 2012.

[17] B. Bross, J. Chen, J.-R. Ohm, GJ Sullivan và Y.-K. Wang, "Sự phát triển trong chuẩn hóa mã hóa video quốc tế sau avc, với tổng quan về mã hóa video đa năng (vvc)," Biên bản báo cáo của IEEE, tập 109, số 9, trang 1463–1493, 2021.

Bài viết này đã được chấp nhận để xuất bản trong IEEE/ACM Transactions on Audio, Speech and Language Processing. Đây là phiên bản của tác giả chưa được biên tập đầy đủ và nội dung có thể thay đổi trước khi xuất bản cuối cùng. Thông tin trích dẫn: DOI 10.1109/TASLP.2023.3277693

[18] BS Atal, “Mã hóa dự đoán giọng nói ở tốc độ bit thấp,” IEEE Transactions on Communications, tập 30, trang 600–614, 1982.

[19] P. Cummiskey, NS Jayant và JL Flanagan, “Lượng tử hóa thích ứng trong mã hóa pcm vi sai của giọng nói,” Tạp chí Kỹ thuật Hệ thống Bell, tập 52, trang 1105–1118, 1973.

[20] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai và Z. Gao, “DVC: Một khuôn khổ nén video sâu từ đầu đến cuối,” trong CVPR, 2019.

[21] J. Li, B. Li và Y. Lu, “Nén video theo ngữ cảnh sâu”, trong NIPS, 2021.

[22] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior và K. Kavukcuoglu, “Wavenet: Một mô hình tạo ra âm thanh thô,” trong Hội thảo tổng hợp giọng nói ISCA lần thứ 9, trang 125–125.

[23] VJM và SJ, “LPCNet: cải thiện tổng hợp giọng nói thần kinh thông qua dự đoán tuyến tính,” trong ICASSP. IEEE, 2019.

[24] A. Spanias, “Mã hóa giọng nói: đánh giá hướng dẫn,” Biên bản của IEEE, tập 82, số 10, trang 1541–1582, 1994.

[25] H. Yang, W. Lim và M. Kim, “Dự đoán đặc điểm thần kinh và mã hóa dư lượng phân biệt gốc cho mã hóa giọng nói tốc độ bit thấp,” trong ICASSP 2023 - Hội nghị quốc tế IEEE về âm học, giọng nói và xử lý tín hiệu (ICASSP), 2023, trang 1–5.

[26] R. Lotfidereshgi và P. Gournay, “Mã hóa nhận thức của lời nói,” trong ICASSP 2022-2022 Hội nghị quốc tế IEEE về âm học, lời nói và xử lý tín hiệu (ICASSP). IEEE, 2022, tr. 7772–7776.

[27] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville và Y. Bengio, “SampleRNN: mô hình tạo âm thanh thần kinh đầu cuối không điều kiện,” trong ICLR, 2017.

[28] J.-M. Valin, K. Vos và TB Terriberry, “Định nghĩa về Opus Audio Codec,” RFC 6716, tháng 9 năm 2012. [Trực tuyến]. Có sẵn: https://www.rfc-editor.org/info/rfc6716

[29] M. Schroeder và B. Atal, “Dự đoán tuyến tính kích thích mã (celp): Giọng nói chất lượng cao ở tốc độ bit rất thấp,” trong ICASSP'85. Hội nghị quốc tế IEEE về âm học, giọng nói và xử lý tín hiệu, tập 10. IEEE, 1985, trang 937–940.

[30] H. Zhou, A. Baevski và M. Auli, “So sánh các mô hình biến tiềm ẩn rời rạc để học biểu diễn giọng nói,” trong IEEE Int. 2021.
Hội nghị về Xử lý tín hiệu giọng nói âm thanh (ICASSP). IEEE, 2021.

[31] A. Baevski, S. Schneider, và M. Auli, “vq-wav2vec: Học tự giám sát các biểu diễn lời nói rời rạc,” trong Hội nghị quốc tế về biểu diễn học tập, 2020. [Trực tuyến]. Có sẵn: https: //openreview.net/forum?id=rylwJxrYDS

[32] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini và LV Gool, “Lượng tử hóa vectơ mềm sang cứng cho các biểu diễn nén học tập đầu cuối đến đầu cuối,” NIPS, 2017.

[33] A. Pandey và D. Wang, “TCNN: Mạng nơ-ron tích chập thời gian để tăng cường giọng nói thời gian thực trong miền thời gian,” trong Hội nghị quốc tế IEEE năm 2019 về Xử lý tín hiệu giọng nói âm thanh (ICASSP). IEEE, 2019.

[34] K. He, X. Zhang, S. Ren và J. Sun, “Đi sâu vào bộ chỉnh lưu: Vượt qua hiệu suất ở cấp độ con người về phân loại imagenet,” trong Biên bản báo cáo hội nghị quốc tế IEEE về thị giác máy tính, 2015, trang 1026–1034.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, AN Gomez, Ł. Kaiser và I. Polosukhin, “Tất cả những gì bạn cần là sự chú ý,” trong Những tiến bộ trong hệ thống xử lý thông tin thần kinh, 2017, trang 5998–6008.

[36] A. Defossez, J. Copet, G. Synnaeve và Y. Adi, “Nén âm thanh thần kinh độ trung thực cao”, bản in trước arXiv arXiv:2210.13438, 2022.

[37] J. Kong, J. Kim và J. Bae, “Hifi-GAN: Mạng đối nghịch tạo ra để tổng hợp giọng nói hiệu quả và có độ trung thực cao,” Những tiến bộ trong Hệ thống xử lý thông tin thần kinh, tập 33, trang 17 022–17 033, 2020.

[38] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, WZ Teoh, J. Sotelo, A. de Brebisson, Y. Bengio và AC Courville, “MelGAN: Mạng đối nghịch tạo sinh để tổng hợp dạng sóng có điều kiện,” Những tiến bộ trong hệ thống xử lý thông tin thần kinh, tập 32, 2019.

[39] B. Xu, N. Wang, T. Chen và M. Li, “Đánh giá thực nghiệm về các hoạt động chỉnh lưu trong mạng tích chập,” bản in trước arXiv arXiv:1505.00853, 2015.

[40] X. Mao, Q. Li, H. Xie, RY Lau, Z. Wang và S. Paul Smolley, “Mạng đối kháng tạo ra theo phương pháp bình phương nhỏ nhất”, trong Biên bản báo cáo hội nghị quốc tế IEEE về thị giác máy tính, 2017, trang 2794–2802.

[41] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, WT Freeman và M. Rubinstein, “Nhìn vào việc lắng nghe tại bữa tiệc cocktail: mô hình nghe nhìn độc lập với người nói để tách giọng nói”, ACM Transactions on Graphics (TOG), tập 37, số 4, trang 1–11, 2018.

[42] S. Wisdom, JR Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton và RA Saurous, “Những ràng buộc về tính nhất quán có thể phân biệt được để cải thiện khả năng nâng cao giọng nói sâu”, trong ICASSP. IEEE, 2019, tr. 900–904.

[43] A. Gritsenko, T. Salimans, R. van den Berg, J. Snoek và N. Kalchbren-ner, “Khoảng cách năng lượng phổ cho tổng hợp giọng nói song song,” Những tiến bộ trong Hệ thống xử lý thông tin thần kinh, tập 33, trang 13 062–13 072, 2020.

[44] M. Dietz, M. Multrus, V. Eksler, V. Malenovsky, E. Norvell, H. Pobloth, L. Miao, Z. Wang, L. Laaksonen, A. Vasilache và cộng sự, “Tổng quan về kiến trúc codec evs,” trong Hội nghị quốc tế về âm học, giọng nói và xử lý tín hiệu (ICASSP) của IEEE năm 2015. IEEE, 2015, tr. 5698–5702.

[45] CKA Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner và S. Srinivasan, “Thử thách giảm tiếng ồn sâu Icassp 2021,” trong ICASSP 2021 - Hội nghị quốc tế IEEE về âm học, giọng nói và xử lý tín hiệu (ICASSP), 2021, trang 6623–6627.

[46] DP Kingma và J. Ba, “Adam: Một phương pháp tối ưu hóa ngẫu nhiên,” trong Hội nghị quốc tế lần thứ 3 về Biểu diễn học tập, ICLR 2015, San Diego, CA, Hoa Kỳ, ngày 7-9 tháng 5 năm 2015, Biên bản báo cáo theo dõi hội nghị, Y. Bengio và Y. LeCun, Biên tập viên, 2015. [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1412.6980 [47] B. Series,
“Phương pháp đánh giá chủ quan mức chất lượng trung gian của hệ thống âm thanh,” Đại hội vô tuyến của Liên minh viễn thông quốc tế, 2014.

[48] I. Rec, “P.862.2: Mở rộng băng thông rộng cho khuyến nghị p.862 để đánh giá mạng điện thoại băng thông rộng và bộ giải mã giọng nói,”
Liên minh Viễn thông Quốc tế, CH-Geneva, 2005.

[49] CHTaal, RCHendriks, R.Heusdens và J.Jensen, “Một biện pháp hiểu khách quan trong thời gian ngắn cho giọng nói ồn ào có trọng số theo thời gian-tần số,” trong ICASSP, 2010.

[50] A. Hines, J. Skoglund, A. Kokaram và N. Harte, “Visqol: Người nghe khách quan về chất lượng giọng nói ảo,” trong IWAENC 2012; Hội thảo quốc tế về tăng cường tín hiệu âm thanh. VDE, 2012, tr. 1–4.

[51] L. Van der Maaten và G. Hinton, “Hình dung dữ liệu bằng t-sne.” Tạp chí nghiên cứu máy học, tập 9, số 11, 2008.

[52] V. Panayotov, G. Chen, D. Povey, và S. Khudanpur, “Librispeech: một kho dữ liệu ASR dựa trên sách nói thuộc phạm vi công cộng,” trong hội nghị quốc tế IEEE năm 2015 về âm học, giọng nói và xử lý tín hiệu (ICASSP). IEEE, 2015, tr. 5206–5210.

[53] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner và M. Sonderegger, “Montreal forcing aligner: Trainingable text-speech alignment using kaldi.” trong Interspeech, tập 2017, 2017, trang 498–502.

[54] K. Ito và L. Johnson, “Bộ dữ liệu lời nói lj,” https://keithito.com/ LJ-Speech-Dataset/, 2017.

[55] SV Kuyk, WB Kleijn và RC Hendriks, “Về tốc độ thông tin của giao tiếp bằng lời nói,” trong ICASSP, 2017, trang 5625–5629.

[56] H. Xue, X. Peng, X. Jiang và Y. Lu, “Hướng tới mã hóa giọng nói thần kinh có khả năng chống lỗi”, trong Proc. Interspeech 2022, 2022, trang 4217–4221.

[57] L. Diener, S. Sootla, S. Branets, A. Saabas, R. Aichner và R. Cutler, “Thử thách che giấu mất gói âm thanh sâu của INTERSpeech 2022,” trong Proc. Interspeech 2022, 2022, trang 580–584.