Saishu và cộng sự. Tạp chí EURASIP về Âm thanh, Lời nói và Âm nhạc
Đang xử lý (2021) 2021:9
https://doi.org/10.1186/s13636-021-00198-4

EURASIP Journal on Audio,
Speech, and Music Processing

NGHIÊN CỨU Truy cập mở

Phư ơ ng pháp nhận dạng dựa trên CNN sự suy giảm tín hiệu tiếng nói

Yuki Saishu1, Amir Hossein Poorjam1,2* và Mads Græsbøll Christensen1

trừu tư ợng

Sự hiện diện của sự suy giảm tín hiệu giọng nói, gây ra sự không khớp âm thanh giữa quá trình đào tạo và vận hành
điều kiện, làm suy giảm hiệu suất của nhiều hệ thống dựa trên giọng nói. Một loạt các kỹ thuật nâng cao có
đư ợc phát triển để bù đắp sự không phù hợp về âm thanh trong các ứng dụng dựa trên giọng nói. Để áp dụng các tín hiệu này
Tuy nhiên, các kỹ thuật nâng cao cần phải biết thông tin trư ớc về sự hiện diện và loại
sự suy giảm tín hiệu tiếng nói. Trong bài báo này, chúng tôi đề xuất một phư ơ ng pháp tiếp cận dựa trên mạng thần kinh tích chập (CNN) mới
để tự động xác định các loại suy giảm chính thư ờng gặp trong các ứng dụng dựa trên giọng nói,
cụ thể là nhiễu cộng, biến dạng phi tuyến và âm vang. Trong phư ơ ng pháp này, một tập hợp các CNN song song, mỗi CNN phát hiện
một loại suy giảm nhất định, đư ợc áp dụng cho biểu đồ phổ log-mel của tín hiệu âm thanh. Kết quả thí nghiệm sử dụng hai
các loại giọng nói khác nhau, cụ thể là giọng nói bệnh lý và lời nói chạy bình thư ờng, cho thấy hiệu quả của
phư ơ ng pháp đư ợc đề xuất trong việc phát hiện sự hiện diện và loại suy giảm tín hiệu giọng nói tốt hơ n
phư ơ ng pháp tiên tiến nhất. Bằng cách sử dụng ánh xạ kích hoạt lớp theo trọng số điểm, chúng tôi cung cấp phân tích trực quan về cách thức
mạng đư a ra quyết định xác định các loại suy giảm khác nhau trong tín hiệu giọng nói bằng cách làm nổi bật các vùng
phổ log-mel có ảnh hư ởng nhiều hơ n đến sự phân hủy mục tiêu.

Từ khóa: Tăng cư ờng tín hiệu, Mạng nơ ron tích chập, Xác định suy thoái, Kiểm soát chất lư ợng,
Hình dung

1. Giới thiệu

Những tiến bộ trong các thiết bị di động như điện thoại thông minh và máy tính bảng đư ợc trang bị micro chất lư ợng cao,
tạo điều kiện thuận lợi cho việc thu thập và xử lý tín hiệu giọng nói trong nhiều môi trư ờng. Tuy nhiên, chất lư ợng của
các bản ghi âm không nhất thiết phải như mong đợi vì chúng có thể có thể bị suy thoái. Trong thực tế, sự có mặt của
suy thoái trong thời gian hoạt động có thể xấu đi
hiệu suất của các hệ thống dựa trên giọng nói, chẳng hạn như lời nói nhận dạng [1], nhận dạng ngư ời nói [2] và phân tích giọng nói bệnh lý (đánh giá tín hiệu giọng nói của ngư ời nói)
bị rối loạn giọng nói) [3, 4], chủ yếu là do âm thanh không khớp giữa điều kiện luyện tập và vận hành. Các
loại suy thoái phổ biến nhất thư ờng gặp trong các ứng dụng dựa trên giọng nói là nhiễu nền,
âm vang và biến dạng phi tuyến.

Tín hiệu giọng nói bị suy giảm do nhiễu cộng, âm vang và biến dạng phi tuyến tư ơ ng ứng có thể đư ợc mô hình hóa như sau:

$$x_n(t) = s(t) + e(t), \quad (1)$$

$$x_r(t) = s(t) \quad h(t), \quad (2)$$

$$x_d(t) = \psi(s(t)), \quad (3)$$

Trong đó t là chỉ số thời gian, s(t) là tín hiệu giọng nói rõ ràng đư ợc ghi bởi micrô trong môi trư ờng không có tiếng ồn và không dội lại, e(t) là tiếng ồn bổ sung, ψ
đại diện cho một hàm phi tuyến, h(t) là xung phòng
phản hồi (RIR) và biểu thị thao tác tích chập. Chúng tôi lư u ý rằng trong thực tế, những sự xuống cấp này thậm chí còn

*Thư từ: ahp@create.aau.dk 1Phòng thí
nghiệm phân tích âm thanh, CREATE, Đại học Aalborg, Rendsburggade 14, 9000 Aalborg, Đan Mạch
2Verisk Analytics, 388 Market Street, 94111 San Francisco, CA, Hoa Kỳ

phức tạp hơ n. Ví dụ, chúng có thể phụ thuộc vào thời gian.
Một loạt các kỹ thuật tăng cư ờng tín hiệu hiệu quả có
đư ợc phát triển để tăng cư ờng tín hiệu giọng nói bị suy giảm
chẳng hạn như giảm tiếng ồn [5, 6], giảm âm vang [7, 8],
và phục hồi một số dạng biến dạng phi tuyến [9,
10]. Hầu hết các thuật toán nâng cao này đã đư ợc
đư ợc thiết kế để giải quyết một loại suy thoái cụ thể trong một
tín hiệu, mặc dù nghiên cứu gần đây về lời nói toàn diện
việc cải tiến, xử lý cả tiếng ồn bổ sung và tiếng vang là đầy hứa hẹn
[11–13]. Tuy nhiên, để đúng cách
bù đắp cho những ảnh hư ởng của sự xuống cấp, cần thiết
để biết hoặc có đư ợc thông tin về sự hiện diện và
loại suy giảm tín hiệu tiếng nói. Kể từ khi hư ởng dẫn sử dụng
Việc kiểm tra tín hiệu rất tốn thời gian, tốn kém,
và thậm chí là không thể trong nhiều ứng dụng dựa trên giọng nói,
một hệ thống phát hiện sự xuống cấp chính xác sẽ rất hữu ích để tự
động xác định sự hiện diện và loại
sự xuống cấp.

Có nhiều cách tiếp cận khác nhau để xác định các
các dạng suy giảm tín hiệu tiếng nói Ví dụ như Mẹ
et al. trong [14] đã đề xuất một phư ơ ng pháp ẩn dựa trên mô hình Markov
Phư ơ ng pháp phân biệt các loại tiếng ồn khác nhau trong lời nói
tín hiệu. Trong một nghiên cứu khác của Desmond et al. [15], cái
tín hiệu dội lại đư ợc phát hiện bằng cách sử dụng một kênh cụ thể
mô hình thống kê Trong [16, 17], việc cắt tín hiệu giọng nói, như
một ví dụ về biến dạng phi tuyến đư ợc phát hiện. Mặc dù
hiệu quả, các phư ơ ng pháp này tập trung vào việc phát hiện một
loại suy thoái duy nhất, cụ thể. Mặt khác, việc sử dụng phân loại
nhiều lớp có thể đư ợc sử dụng để
phát hiện các loại suy thoái khác nhau. Trong [18, 19], Poorjam
et al. đề xuất hai phư ơ ng pháp tiếp cận dựa trên phân loại đa lớp
tổng quát phát hiện các loại suy thoái khác nhau,
chỉ điều tra về tín hiệu giọng nói bệnh lý và
độ chính xác vẫn còn chư a đầy đủ. Hơ n nữa, không có
kiểm soát việc phân công lớp học theo những cách tiếp cận này
khi một loại suy thoái mới đư ợc quan sát thấy
bộ phân loại chư a đư ợc huấn luyện. Ví dụ: clip-ping, mất gói, nén
phạm vi động, tự động
giành quyền kiểm soát và biến dạng do sử dụng chất lư ợng thấp hoặc
thiết bị có cấu hình không đúng đư ợc coi là mới
các loại suy thoái đối với bộ phân loại nhiều lớp chỉ đư ợc đào tạo
với tín hiệu ồn ào và vang dội.

Để khắc phục những hạn chế của mô hình dựa trên nhiều lớp
tiếp cận, ngư ời ta có thể sử dụng một cách phân loại đa dạng
cách tiếp cận trong đó có thể có nhiều nhãn lớp
đư ợc ấn định cho từng mẫu. So với các phư ơ ng pháp dựa trên nhiều
lớp, cách tiếp cận này có thể giải quyết tốt hơ n một số
những trư ờng hợp khó khăn như sự xuất hiện của một loại suy thoái
mới và khi có nhiều hơ n một loại suy thoái cùng tồn tại. Trong
trư ờng hợp trư ớc, mẫu có thể đư ợc phân loại
như không có lớp mục tiêu nào. Trong trư ờng hợp sau, hơ n
nhiều hơ n một máy dò có thể chấp nhận một tín hiệu có sự suy giảm
hỗn hợp. Một giải pháp khả thi là tích hợp các thuật toán hiện có,
đư ợc phát triển để phát hiện từng
loại suy thoái, thành một khuôn khổ thống nhất và coi mỗi hệ thống
con như một máy dò để đư a ra quyết định
về một tín hiệu. Tuy nhiên, các thuật toán đư ợc phát triển độc lập
có thể đư a ra những giả định rất khác nhau.
và có thể có những yêu cầu đa dạng mà đôi khi có thể xung đột. Vì
vậy, việc tích hợp chúng thành một
khuôn khổ này rất khó khăn và việc đáp ứng tất cả các yêu cầu cùng
một lúc có thể không khả thi ở một số nơ i.
các trư ờng hợp.

Là một giải pháp thay thế, Poorjam et al. đề xuất một
phư ơ ng pháp tiếp cận dựa trên dữ liệu sử dụng một tập hợp các mô
hình hỗn hợp Gaus-sian song song (GMM) để phát hiện ba loại
suy giảm tín hiệu giọng nói bệnh lý, cụ thể là nhiễu nền, âm vang và
biến dạng phi tuyến [4].
Tất cả các máy dò trong phư ơ ng pháp này đều giống nhau về mặt
phức tạp, các giả định cơ bản và các tính năng âm thanh ngoại trừ
việc chúng đư ợc huấn luyện bằng cách sử dụng các mức suy giảm khác nhau
tín hiệu. Cách tiếp cận này tập trung vào những tiếng nói bệnh lý
và đặc biệt là các nguyên âm ngân.

Trong bài báo này, chúng tôi đề xuất một cách tiếp cận dựa trên
mạng thần kinh tích chập (CNN) chính xác hơ n có thể
xác định sự xuống cấp không chỉ ở các nguyên âm duy trì, mà còn
cũng như trong lời nói chạy bình thư ờng. CNN là mạng lư ới thần kinh
sâu hiệu quả về mặt tính toán, có khả năng học hỏi
các mẫu phức tạp trong biểu đồ phổ của tín hiệu giọng nói.
Theo cách tiếp cận này, chúng tôi áp dụng một tập hợp các CNN song song cho
phổ log-mel của tín hiệu. Mỗi mô hình CNN,
đư ợc huấn luyện với các tín hiệu bị hỏng do sự suy giảm cụ thể
chịu trách nhiệm phát hiện sự suy giảm tư ơ ng ứng trong tín hiệu kiểm
tra. Điểm dự đoán của một điều không thể nhìn thấy
mẫu thử nghiệm có thể đư ợc sử dụng để liên kết nhiều nhãn phân hủy
với một quan sát và có thể đư ợc hiểu là
mức độ đóng góp của mỗi sự suy thoái trong một
tín hiệu. Hơ n nữa, bằng cách sử dụng ánh xạ kích hoạt lớp điểm
(score-CAM) [20], chúng tôi giải thích một cách trực quan về những gì
cơ sở, các mô hình CNN đư a ra quyết định cụ thể trong việc phát hiện
các loại suy thoái khác nhau bằng cách tìm các vùng
trong biểu đồ phổ thang đo mel của tín hiệu bị suy giảm
có ảnh hư ởng lớn nhất đến điểm số của lớp mục tiêu. TRONG
kỹ thuật này, các bản đồ kích hoạt khác nhau đư ợc áp dụng cho
phổ đầu vào, mỗi phổ gây nhiễu loạn một vùng của
quang phổ. Sau đó, ảnh hư ởng của từng bản đồ kích hoạt lên
điểm dự đoán đư ợc quan sát. Tầm quan trọng của mỗi
bản đồ kích hoạt đư ợc xác định bởi điểm dự đoán trên
lớp mục tiêu. Cuối cùng, một bản đồ vị trí nổi bật đư ợc tạo ra bởi một
sự kết hợp tuyến tính có trọng số của tất cả các bản đồ kích hoạt để
trực quan hóa biểu diễn bên trong trong CNN [20]. Vì điều này
Kỹ thuật này không yêu cầu bất kỳ sửa đổi nào đối với kiến trúc của
mạng, nó có thể đư ợc áp dụng cho nhiều loại
của các mô hình CNN.

Phần còn lại của bài viết này đư ợc tổ chức như sau. TRONG
Phần 2, chúng tôi xây dựng bài toán phát hiện sự suy giảm tự động và
mô tả phư ơ ng pháp đề xuất.
Bố trí thử nghiệm đư ợc giải thích ở Phần 3. Trong

Phần 4, chúng tôi trình bày và thảo luận về kết quả. Bài viết kết thúc với kết luận ở Phần 5.

2 Mô tả hệ thống 2.1 Xây dựng bài toán Trong bài

toán phát hiện suy giảm chất lượng tín hiệu giọng nói, chúng ta được cấp một tập dữ liệu huấn luyện = {xn, yn,d} trong $_{n=1}^{N}$, đó xn Rk biểu thị quan sát thứ n của k chiều. Tùy thuộc vào hệ thống và mức độ xử lý, điều này có thể biểu thị các đặc điểm âm thanh của tín hiệu âm thanh hoặc khung tín hiệu. Ví dụ: trong hệ thống đề xuất, được giới thiệu ở Phần 2.2, xn biểu thị phổ log-mel của tín hiệu âm thanh thứ n và trong hệ thống cơ sở, được mô tả ở Phần 2.3, nó là hệ số mel-tần số của tín hiệu âm thanh thứ n. khung tín hiệu yn,d {0, 1} biểu thị liệu quan sát thứ n có thuộc lớp suy biến d hay không. N là tổng số mẫu huấn luyện. Mục tiêu là ước chừng hàm phân loại nhị phân gd cho từng loại suy giảm d, sao cho đối với một quan sát không có trong dữ liệu huấn luyện, xtest, xác suất của mẫu thử nghiệm được phân loại vào đúng lớp là tối đa. Nói cách khác, nhãn suy giảm ước tính, yˆd = gd(xtest), với yˆd [0, 1], càng gần với nhãn thực càng tốt.

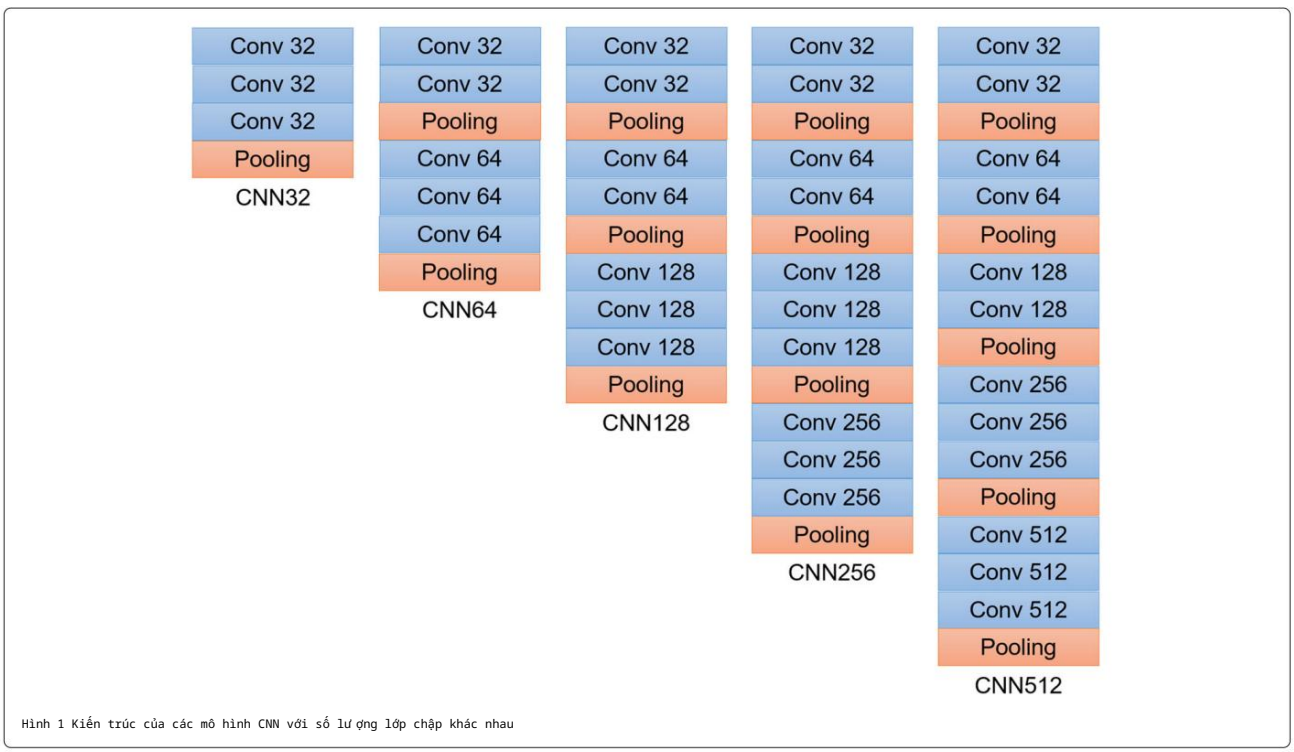
2.2 Phương pháp đề xuất

Trong phương pháp đề xuất, chúng tôi sử dụng một tập các CNN song song để tính gần đúng các hàm gd. Mỗi CNN, lấy cảm hứng từ VGGNet [21], bao gồm một số khối tích chập, và mỗi khối bao gồm một số lớp chập với kích thước hạt nhân 3 × 3. Như được hiển thị trong Hình 1, chúng tôi đề xuất 5 kiến trúc CNN khác nhau cho mỗi máy dò để nghiên cứu kiến trúc tối ưu cho vấn đề phát hiện suy thoái. Các con số phía trước "Conv" trong mỗi lớp hiển thị số lượng bản đồ tính năng. CNN32, có 28.807 tham số cần huấn luyện, bao gồm một khối tích chập gồm 3 lớp. CNN64, với 120.423 tham số, bao gồm khối chập 2 lớp và 3 lớp.

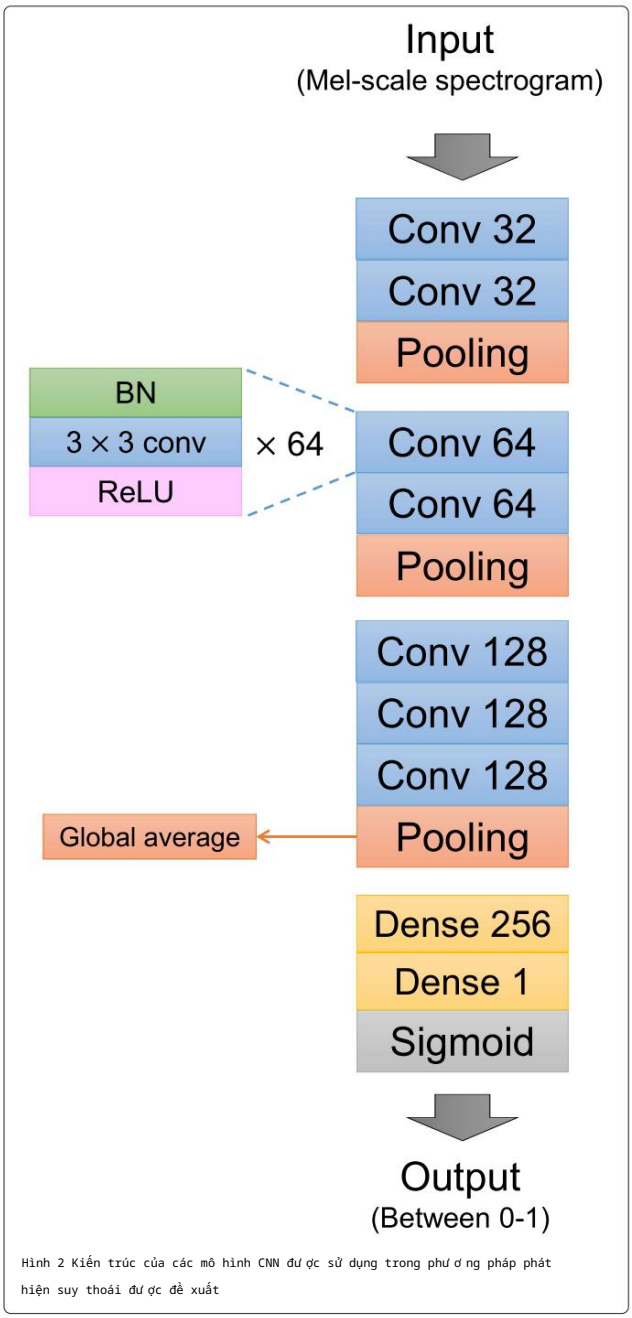
CNN128 bao gồm hai khối tích chập 2 lớp và một khối 3 lớp. Số lượng tham số của mạng này là 469.543. Trong CNN256 có hai khối 2 lớp và hai khối 3 lớp và có 1.979.175 tham số. Cuối cùng, CNN512 bao gồm 7.947.559 tham số, được tạo thành từ ba khối 2 lớp và hai khối 3 lớp. Trong Hình 2, kiến trúc của CNN128 được minh họa chi tiết hơn.

Để kết nối các lớp tích chập, chúng tôi sử dụng chuẩn hóa hàng loạt (BN) và đơn vị tuyến tính chỉnh lưu (ReLU). BN cho phép một mạng lưới thần kinh sâu học với tốc độ học lớn hơn, tạo điều kiện cho sự hội tụ nhanh hơn và khái quát hóa tốt hơn [22]. Lớp đầu ra bao gồm hai lớp dày đặc—còn được gọi là các lớp được kết nối đầy đủ—được kết nối với lớp tích chập cuối cùng bằng cách gộp tuổi trung bình toàn cầu. Chúng tôi sử dụng hàm kích hoạt sigmoid trong lớp đầu ra để tạo ra điểm trong phạm vi [0, 1].

Là tính năng âm thanh, chúng tôi sử dụng biểu đồ phổ log-mel có kích thước 300 khung hình × thùng 40 mel, được tính bằng cách lấy logarit của đầu ra của dãy bộ lọc thang đo mel được áp dụng cho biến đổi Fourier thời gian ngắn (STFT) của tín hiệu .



Hình 1 Kiến trúc của các mô hình CNN với số lượng lớp chập khác nhau



Hình 2 Kiến trúc của các mô hình CNN đư ợc sử dụng trong phư ơ ng pháp phát hiện suy thoái đư ợc đề xuất

Quang phổ log-mel là một kỹ thuật tham số hóa tín hiệu phổ biến trong nhiều ứng dụng âm thanh sử dụng mạng thần kinh sâu, cung cấp biểu diễn 2 chiều hiệu quả, phù hợp về mặt nhận thức của tín hiệu âm thanh.
So với STFT, biểu đồ phổ log-mel cung cấp biểu diễn tín hiệu âm thanh ít dư thừa hơ n và cho phép CNN học với số lư ợng dữ liệu huấn luyện ít hơ n. Thang đo decibel đư ợc thúc đẩy bởi nhận thức của con ngư ời về âm lư ợng [23] và đã cho thấy khả năng phân biệt tốt hơ n so với phiên bản tuyến tính [24]. Biểu đồ phổ log-mel thu đư ợc cùng với biểu đồ đầu tiên và đạo hàm bậc hai đư ợc sử dụng làm đặc tính đầu vào cho CNN.

Chúng tôi sử dụng phư ơ ng pháp giảm độ dốc ngẫu nhiên (SGD) để giảm thiểu entropy chéo nhị phân cho mỗi phân loại đư ợc xác định bằng:

$$L = \frac{1}{N}\sum_{n=1}^{N} y_{n,d}\ln(g_d(x_n)) + (1 - y_{n,d})\ln(1 - g_d(x_n)), \quad (4)$$

trong đó $g_d(x_n) \in [0, 1]$ là điểm đầu ra của CNN đư ợc đào tạo để xác định một loại suy thoái cụ thể và $y_{n,d} \in \{0, 1\}$ là nhãn suy giảm thực sự.

Quyết định quan sát thử nghiệm đư ợc đư a ra bằng cách đặt ngư ỡng cho điểm đầu ra của mỗi CNN. Bằng cách này, nếu một mẫu thử nghiệm phải chịu một kiểu phân hủy mới, chúng tôi hy vọng nó sẽ bị tất cả các CNN từ chối dựa trên ngư ỡng đư ợc xác định trư ớc. Hơ n nữa, nếu một quan sát bị ảnh hư ởng bởi nhiều hơ n một loại suy thoái, chúng tôi hy vọng rằng điểm đầu ra của nhiều CNN sẽ vư ợt quá ngư ỡng. Cần lư u ý rằng do việc lựa chọn ngư ỡng quyết định tối ư u phụ thuộc vào ứng dụng nên trong nghiên cứu này, chúng tôi xem xét điểm mềm và sử dụng thư ớc đo độc lập với ngư ỡng, đư ợc giới thiệu trong Phần 3.4, để đánh giá hiệu suất của hệ thống đư ợc đề xuất .

2.3 Hệ thống cơ sở

Là một hệ thống cơ sở mà chúng tôi so sánh với hệ thống đư ợc đề xuất của mình, chúng tôi sử dụng phư ơ ng pháp phát hiện sự suy giảm mô hình nền phổ quát mô hình hỗn hợp Gaussian (GMM-UBM) đư ợc đề xuất trong [4] . Theo cách tiếp cận này, một tập hợp các GMM song song, đư ợc gắn vào các khung của tín hiệu giọng nói trong miền hệ số cao tần số mel (MFCC), đư ợc sử dụng để phát hiện các loại suy giảm khác nhau. Giai đoạn huấn luyện bao gồm hai bư ớc: (1) huấn luyện GMM không phụ thuộc vào suy thoái với một lư ợng lớn dữ liệu huấn luyện từ các lớp suy thoái khác nhau, đư ợc gọi là UBM và (2) huấn luyện một tập hợp GMM phụ thuộc vào suy thoái bằng cách điều chỉnh các tham số của UBM sử dụng dữ liệu huấn luyện tư ơ ng ứng. Để đánh giá, điểm xác định của một loại suy giảm nhất định, d và đầu vào trình tự thời gian, $X = (x_1, \ldots, x_n, \ldots, x_N)$, đư ợc tính theo phư ơ ng trình sau:

$$\sigma_d = g_d(X) = \frac{1}{N}\sum_{n=1}^{N}\log p(x_n|\lambda_d) - \sum_{n=1}^{N}\log p(x_n|\lambda_{ubm}), \quad (5)$$

Trong đó N là tổng số khung thời gian, λ_{ubm} và λ_d lần lư ợt là các tham số của UBM và GMM phụ thuộc vào sự suy giảm và $p(x_n|\lambda)$ là hàm mật độ xác suất Gaussian. Việc xác định đư ợc thực hiện bằng cách đặt ngư ỡng cho điểm số.

3 Bố trí thí nghiệm

3.1 Bộ dữ liệu

Cách tiếp cận của chúng tôi có thể được áp dụng cho bất kỳ loại lời nói nào, chẳng hạn như lời nói chạy bình thường, lời nói thì thầm, lời nói giàu cảm xúc, phát âm nguyên âm kéo dài và giọng hát. Trong nghiên cứu này, chúng tôi xem xét hai loại giọng nói, đó là giọng nói bệnh lý và lời nói chạy bình thường, để đánh giá hiệu quả của phương pháp được đề xuất. Đối với giọng nói bệnh lý, chúng tôi đã sử dụng bộ dữ liệu bệnh Parkinson di động (MMPD) mPower [25] bao gồm hơn 65.000 mẫu giọng nói gồm các âm vị duy trì trong 10 giây của nguyên âm /a/ được ghi ở tần số lấy mẫu 44,1 kHz bởi bệnh nhân PD và loa khỏe. Tập dữ liệu này đã được chọn vì hầu hết bệnh nhân PD đều mắc một số dạng rối loạn giọng nói [26]. Hơn nữa, vì các âm vị nguyên âm duy trì cung cấp một cấu trúc âm thanh đơn giản để mô tả nguồn thanh hầu và cấu trúc cộng hưởng của đường phát âm [27], nên chúng được coi là tài liệu giọng nói chính để phân tích giọng nói bệnh lý do một loạt các rối loạn y tế gây ra. Đối với giọng nói thông thường, chúng tôi sử dụng cơ sở dữ liệu giọng nói tiếng Anh do Trung tâm Nghiên cứu Công nghệ Giọng nói tại Đại học Edinburgh xuất bản [28]. Các mẫu của cơ sở dữ liệu này được ghi ở tần số 48 kHz.

3.1.1 Giọng nói bệnh lý Để

chuẩn bị dữ liệu cho thí nghiệm phát hiện sự suy giảm giọng nói bệnh lý, chúng tôi chọn ngẫu nhiên 9.000 mẫu từ bộ dữ liệu MMPD và chia thành 5 nhóm bằng nhau gồm 1.800 mẫu. Các bản ghi âm của nhóm đầu tiên đã bị suy giảm do sáu loại tiếng ồn phụ gia khác nhau, đó là tiếng ồn lảm nhảm, tiếng đường phố, nhà hàng, văn phòng, tiếng Gauss trắng và tiếng gió1 trong các điều kiện tỷ lệ tín hiệu trên tiếng ồn (SNR) khác nhau, từ 10 dB đến 20 dB. Tín hiệu nhiễu được lấy mẫu lại thành 44,1 kHz trước khi được thêm vào tín hiệu thoại. Để giảm xác suất quan sát các tín hiệu bị suy giảm bởi các phân đoạn nhiễu giống hệt nhau trong cả tập con huấn luyện và đánh giá, chúng tôi đã thêm một phân đoạn ngẫu nhiên của tệp nhiễu vào mỗi tín hiệu sạch.

Các bản ghi của nhóm thứ hai được lọc bằng 46 phản hồi xung trong phòng thực (RIR) của cơ sở dữ liệu AIR [29], được đo bằng một chiếc điện thoại mô phỏng ở các vị trí cầm tay và rảnh tay trong nhiều môi trường trong nhà thực tế khác nhau, chẳng hạn như làm phòng họp, hành lang, giảng đường, văn phòng, cầu thang, bếp để sản xuất các mẫu kiến vang. Thời gian vang của RIR, RT60, được định nghĩa là thời gian để âm thanh đã tắt giảm đi 60 dB [30], nằm trong khoảng từ 390 ms đến 1,47 giây. Các

1Các tệp tiếng ồn ào, nhà hàng và đường phố được lấy từ https://www.soundjay.com, tiếng ồn văn phòng được lấy từ https://freesound.org/people/ DavidFrbr/sounds/327497, tiếng ồn trắng được lấy từ https://www. audiocheck.net/testtones_whitenoise.php, và tiếng ồn của gió được lấy từ https://www.iks.rwth-aachen.de/forschung/tools-downloads/databases/wind-noise-database .

Tỷ lệ năng lượng trực tiếp và năng lượng dội lại của RIR nằm trong khoảng từ 4,35 đến 12,28 dB. Các RIR đã được lấy mẫu lại thành 44,1 kHz trước khi tích chập.

Các mẫu của nhóm thứ ba bị biến dạng bằng cách cắt, mã hóa hoặc cắt, sau đó là mã hóa như một ví dụ về biến dạng phi tuyến. Mức cắt, được định nghĩa là tỷ lệ của biên độ tín hiệu tuyệt đối cực đại mà các giá trị mẫu lớn hơn ngưỡng này bị giới hạn, được đặt thành 0,3, 0,5 hoặc 0,7 và chúng tôi đã sử dụng mã 9,6 kbps và 16 kbps- codec dự đoán tuyến tính kích thích (CELP) [31].

Chúng tôi sử dụng nhóm thứ tư để kết hợp tiếng ồn bổ sung và tiếng vang, trong đó mẫu giọng nói được lọc bởi RIR và được thêm vào tín hiệu nhiễu cũng được tích hợp với RIR. Các tín hiệu nhiễu trong trường hợp này bị suy giảm do tiếng ồn môi trường trong nhà như tiếng ồn lảm nhảm, nhà hàng và văn phòng trong các điều kiện SNR 0 dB, 5 dB hoặc 10 dB. Lý do chọn tập hợp con này là để đánh giá xem một tín hiệu, trong đó cả tiếng ồn và tiếng vang cùng tồn tại, có thể được phát hiện bằng cả máy dò tiếng ồn và tiếng vang hay không. Nhóm thứ năm được sử dụng mà không cần xử lý và được coi là loại sạch.

3.1.2 Giọng nói chạy bình thường

Để chuẩn bị mẫu cho các lớp ồn và ồn- vang trong giọng nói bình thường, chúng tôi đã sử dụng bộ dữ liệu giọng nói song song sạch và ồn (NS) [28] và tập dữ liệu giọng nói sạch và ồn-rộng vang (NRS) [32], tương ứng.

Trong bộ dữ liệu NS, tín hiệu giọng nói rõ ràng, được ghi lại bởi 28 loa cân bằng giới tính, chịu 10 tiếng ồn khác nhau thu được từ cơ sở dữ liệu DEMAND [33] ở mức 0 dB, 5 dB, 10 dB và 15 dB SNR. Từ tập hợp con sạch của tập dữ liệu này, chúng tôi đã chọn ngẫu nhiên 1.800 mẫu cho lớp sạch và 1.800 mẫu không chồng chéo từ tập hợp con ồn ào cho lớp ồn ào.

Trong cơ sở dữ liệu NRS, giọng nói vang dội ồn ào được tạo ra bằng cách kết hợp tín hiệu sạch với RIR và thêm tín hiệu đó vào tín hiệu nhiễu cũng được tích hợp với phản hồi xung trong phòng. Vì vậy, chúng tôi đã chọn ngẫu nhiên 1800 mẫu cho lớp tiếng ồn vang dội. Để chuẩn bị dữ liệu cho các lớp biến dạng dội lại và phi tuyến, chúng tôi đã chọn hai tập hợp con riêng biệt gồm 1800 mẫu từ phần sạch của tập dữ liệu và phân hủy chúng theo cách tương tự như để tạo các lớp biến dạng dội lại và phi tuyến cho giọng nói bệnh lý.

3.2 Đặc điểm âm thanh

Chúng tôi chuẩn hóa các tín hiệu bằng cách trừ giá trị trung bình và chia cho biên độ cực đại tuyệt đối. Sau đó, đối với đầu vào của CNN, chúng tôi đã phân đoạn tín hiệu thành các khung 30 ms với độ chồng lấp 10 ms bằng cửa sổ Hamming.

Sau đó, đối với mỗi khung hình của tín hiệu, chúng tôi đã tính toán biểu đồ phổ log-mel của 40 kênh cùng với các đạo hàm thứ nhất và thứ hai.

Là đầu vào cho hệ thống GMM-UBM, chúng tôi đã sử dụng MFCC được tính toán bằng cách sử dụng cửa sổ Hamming 30 ms với 10 chồng chéo ms và ngân hàng bộ lọc quy mô mel 27 kênh. Vì mỗi khung của tín hiệu, 13 hệ số, bao gồm log-năng lượng của khung, cùng với hệ số thứ nhất và thứ hai dẫn xuất của MFCC đã được tính toán để tạo thành một Vector đặc trưng 39 chiều. Chúng tôi đã sử dụng các giá trị giống nhau cho các tham số của hệ thống cơ sở như đã được sử dụng trong [4] để tái tạo kết quả của họ.

3.3 Thông số cấu hình

Tất cả các mạng CNN trong thí nghiệm của chúng tôi đều được đào tạo 20 kỷ nguyên bằng cách sử dụng SGD để giảm thiểu hàm mất entropy chéo nhị phân được xác định trong biểu thức. (4). Độ lớn của các dao động ngẫu nhiên trong động lực học SGD được biểu thị bằng thang nhiễu, ρ, tỷ lệ thuận với tốc độ hội tụ và được định nghĩa là [34]:

$$\rho = \frac{N}{B} \frac{1}{1 \quad \nu} \approx \frac{N}{B(1 \quad \nu)}, \quad \text{(NB)}, \qquad (6)$$

Trong đó N là số lượng mẫu đào tạo, là tốc độ học, B là kích thước lô và ν là động lượng của SGD. Trong các thử nghiệm của chúng tôi, kích thước lô trong mỗi kỷ nguyên và động lượng của SGD lần lượt được đặt ở mức 64 và 0,9. Chúng tôi cũng giảm tỷ lệ học tập theo cấp số nhân từ 0,01 đến 0,0001 từ kỷ nguyên này sang kỷ nguyên khác.

Đối với hệ thống cơ sở, số lượng thành phần hỗn hợp nents được đặt thành 1024 theo [4].

3.4 Chỉ số hiệu suất

Để đánh giá hiệu suất của hệ thống được đề xuất, chúng tôi đã sử dụng khu vực dưới đường cong đặc tính vận hành máy thu (ROC) (AUC). Trong đường cong ROC, giá trị thực tỷ lệ dương được vẽ theo tỷ lệ dương tính giả cho ngưỡng quyết định khác nhau của điểm số. AUC tóm tắt đường cong ROC thành một số duy nhất tạo thuận lợi cho so sánh dễ dàng hơn giữa các hệ thống khác nhau bất kể ngưỡng quyết định là ứng dụng và tham số phụ thuộc vào người dùng. Giá trị AUC bằng 0,5 đại diện cho hiệu suất ở cấp độ cơ hội, trong khi AUC bằng 1 có nghĩa là sự phân tách hoàn hảo giữa các lớp.

4 Kết quả và thảo luận

CNN là một máy biến áp phi tuyến phức tạp, có thể cung cấp nhiều biến thể biểu thức của đầu vào thông qua các lớp. Bằng cách tăng số lượng tham số, thường có thể đạt được biểu thức đầu vào tốt hơn ở mức chi phí tăng nguy cơ trang bị quá mức như mô hình có thể ghi nhớ chi tiết cụ thể của dữ liệu đào tạo. Vì vậy, trước tiên chúng tôi tiến hành thử nghiệm để chọn ra phương án tối ưu Kiến trúc CNN cho bài toán phát hiện sự xuống cấp và sử dụng nó cho các thí nghiệm còn lại. Sau đó, sau khi so sánh hiệu quả của phương pháp đề xuất với cơ bản, chúng tôi giải thích trực quan cách CNN đưa ra quyết định để xác định sự suy giảm tín hiệu tiếng nói.

Trong tất cả các thử nghiệm, chúng tôi đã sử dụng xác thực chéo 10 lần (CV) trong đó các mẫu được chia ngẫu nhiên thành 10 tập hợp con không chồng chéo và có kích thước bằng nhau. Sau đó, 9 trên 10 các tập hợp con được sử dụng để huấn luyện các mô hình và tập hợp con còn lại được sử dụng để đánh giá. Thủ tục này đã được lặp lại 10 lần để tất cả các tập hợp con được sử dụng một lần cho huấn luyện và đánh giá mô hình. Cần lưu ý rằng để đánh giá, chúng tôi đã mở rộng từng tập hợp con thử nghiệm bằng cách thêm 20 các mẫu ngoại lệ không chứa thông tin liên quan đến bối cảnh của các tập dữ liệu, chẳng hạn như tiếng chó sủa hoặc bản ghi âm lời thì thầm để thể hiện liệu các máy dò có thể loại bỏ các mẫu ngoại lệ đó hay không.

Để nghiên cứu kiến trúc tốt nhất cho các mô hình CNN, chúng tôi so sánh hiệu suất của CNN32, CNN64, CNN128, CNN256 và CNN512. Những kiến trúc này là được giải thích ở Phần 2.2 và được minh họa ở Hình 1. Trong phần này thí nghiệm, chúng tôi đã sử dụng những giọng điệu bệnh hoạn. Kết quả, được báo cáo trong Bảng 1, cho thấy sự khác biệt về hiệu suất giữa các kiến trúc mạng khác nhau là cận biên, đặc biệt là để phát hiện tiếng ồn trong đó tất cả các công trình mạng đều hoạt động tốt như nhau. Tuy nhiên, việc có một mạng lưới một kiến trúc đơn giản hơn thể hiện hiệu suất cao hơn sẽ được mong muốn hơn để giảm nguy cơ trang bị quá mức. Xem xét số lượng tham số của từng mô hình, được đề cập trong Phần 2.2, và vì CNN128 vượt trội hơn các loại khác trong việc xác định độ méo và tiếng vang và có độ phức tạp và độ chính xác cân bằng nhất cho ứng dụng của mình, chúng tôi chọn kiến trúc này cho các thí nghiệm còn lại

Khi kiến trúc CNN tối ưu được chọn, chúng ta có thể so sánh một cách khách quan hiệu suất của hệ thống được đề xuất với đường cơ sở. Như đã giải thích ở Phần 2.3 , Giai đoạn huấn luyện trong hệ thống cơ bản bao gồm hai bước, cụ thể là đào tạo một UBM với số lượng lớn các khóa đào tạo các mẫu từ các lớp phân hủy khác nhau và điều chỉnh các mô hình phụ thuộc vào sự suy thoái với các mô hình tương ứng mẫu đào tạo. Để đào tạo UBM, chúng tôi đã sử dụng 8000 mẫu (1600 mẫu từ mỗi lớp). Phần còn lại 1000 mẫu (200 mẫu từ mỗi lớp) đã được sử dụng để thích ứng và đánh giá sự suy thoái phụ thuộc GMM. Để đưa ra sự so sánh công bằng giữa phương pháp được đề xuất và hệ thống cơ sở về mặt

Bảng 1 So sánh hiệu suất của các CNN khác nhau kiến trúc trên tập dữ liệu giọng nói bệnh lý ở dạng trung bình AUC±95% khoảng tin cậy

Máy dò	CNN32	CNN64	CNN128	CNN256	CNN512
Tiếng ồn	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00
Méo mó	0,98±0,01	0,98±0,01	0,99±0,00	0,98±0,01	0,98±0,01
Âm vang	0,90±0,01	0,91±0,01	0,93±0,01	0,91±0,01	0,89±0,01

Những con số in đậm thể hiện hiệu suất tốt nhất

Bảng 2 Hiệu suất của hệ thống CNN128 khi không có các thông số đã được chia sẻ giữa các máy dò và khi các thông số ở một số lớp đã được chia sẻ

Máy dò	Đào tạo độc lập (cách tiếp cận đầu tiên)	Chia sẻ thông số (cách tiếp cận thứ hai)
Tiếng ồn	1,00±0,00	1,00±0,00
Méo mó	0,99±0,00	0,98±0,01
Tiếng vang	0,93±0,01	0,88±0,01

Các kết quả ở dạng trung bình khoảng tin cậy AUC±95%

dữ liệu được sử dụng, chúng tôi thực hiện hai cách tiếp cận khác nhau. TRONG Cách tiếp cận đầu tiên, chúng tôi huấn luyện từng bộ phân loại nhị phân từ đầu bằng cách sử dụng tất cả các mẫu đào tạo tương ứng. TRONG Mặt khác, cách tiếp cận thứ hai là chúng tôi đã huấn luyện một bộ phân loại nhiều lớp với các mẫu huấn luyện được sử dụng để huấn luyện mô hình UBM. Sau đó, sử dụng các mẫu đã khai thác để điều chỉnh các GMM phụ thuộc vào sự suy thoái, chúng tôi đã tinh chỉnh ba bộ phân loại nhị phân từ đa lớp được đào tạo bộ phân loại. Ở bước tinh chỉnh ta giữ nguyên các thông số của khối chập thứ nhất và thứ hai bị đóng băng và điều chỉnh các tham số của khối tích chập cuối cùng và các lớp được kết nối đầy đủ. Bằng cách này, tương tự như hệ thống cơ sở, các thông số của hai khối đầu tiên đã được chia sẻ trên mỗi máy dò. Bảng 2 cho thấy hiệu suất của CNN128 trên dữ liệu giọng nói bệnh lý được thiết lập khi hai cách tiếp cận này được áp dụng. Chúng ta có thể quan sát rằng các mô hình, đặc biệt là độ vang máy dò, hoạt động tốt hơn khi các bộ phân loại được đào tạo độc lập. Vì vậy, chúng tôi đã sử dụng phương pháp đầu tiên khi so sánh phương pháp đề xuất của chúng tôi với đường cơ sở hệ thống.

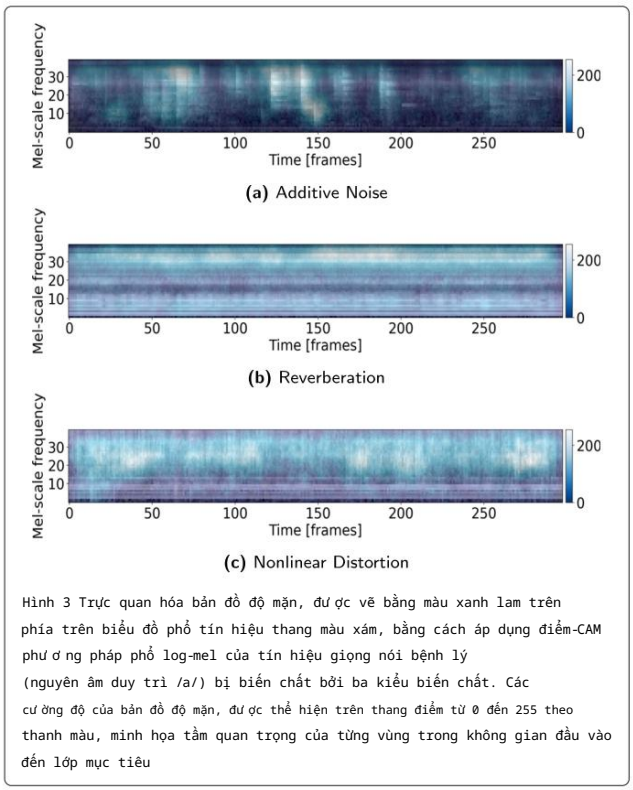
Bảng 3 cho thấy hiệu suất của đường cơ sở và các hệ thống được đề xuất. Kết quả cho thấy đề xuất hệ thống vượt trội hơn so với đường cơ sở cho cả bệnh lý giọng nói và tín hiệu giọng nói đang chạy, đặc biệt để xác định âm vang trong giọng nói bệnh lý và phụ gia tiếng ồn trong lời nói đang chạy. Chúng ta có thể nhận thấy rằng cả hai hệ thống đều có xu hướng chung là hiệu suất của máy dò âm vang thấp hơn nhiều so với tiếng ồn máy dò, chủ yếu là do nhận dạng sai các bản ghi trong đó tiếng ồn và tiếng vang cùng tồn tại, nhưng tiếng ồn chiếm ưu thế hơn. Hơn nữa, kết quả chỉ ra rằng việc xác định âm vang trong giọng nói bệnh lý là thách thức đối với hệ thống cơ sở. Điều này là do không giống như lời nói chạy, sự bao bọc tạm thời của một nguyên âm duy trì không đạt đỉnh và do đó không bị ảnh hưởng nhiều bởi tiếng vang. Hơn nữa, vì đường viền cao độ của một nguyên âm ngân gần như giữ nguyên trong một khoảng thời gian ngắn. về thời gian so với lời nói đang chạy, động những thay đổi trong miền tần số ít bị ảnh hưởng hơn trong nguyên âm duy trì hơn so với lời nói chạy. Những điều này làm cho việc xác định tiếng vang khó khăn hơn đối với hệ thống cơ sở. Tuy nhiên, mô hình CNN có thể tốt hơn phân biệt được những khác biệt tinh tế này.

Bảng 3 So sánh giữa phương pháp đề xuất cho phát hiện sự xuống cấp và hệ thống cơ sở cho bệnh lý giọng nói và lời nói chạy bình thường

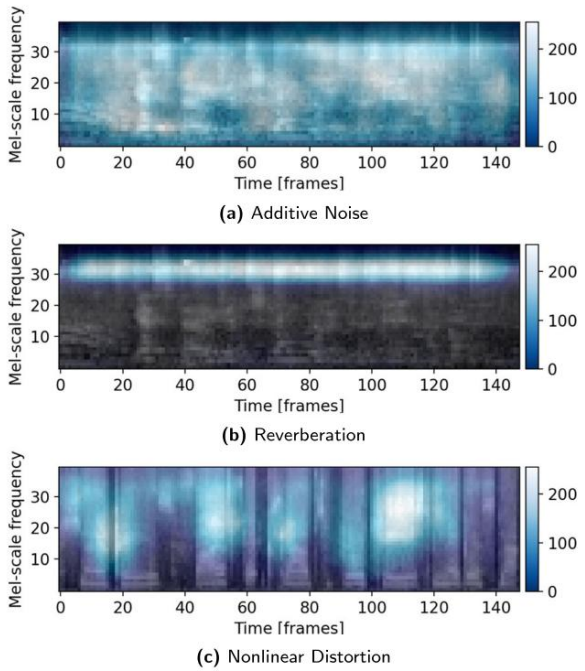
Máy dò	Giọng nói bệnh lý		Lời nói chạy bình thường	
	Đường cơ sở	được đề xuất		Đề xuất
Tiếng ồn	0,96±0,00	1,00±0,00	0,71±0,00	0,95±0,01
Méo mó	0,90±0,01	0,99±0,00	0,83±0,01	1,00±0,00
Âm vang	0,75±0,00	0,93±0,01	0,84±0,01	0,99±0,00

Các kết quả ở dạng khoảng tin cậy AUC trung bình ±95% và số in đậm thể hiện hiệu suất tốt nhất

Mặt khác, vì nội dung tần số và đặc điểm của một số loại tiếng ồn nền, chẳng hạn như như lảm nhảm, tương tự như tín hiệu giọng nói đang chạy, việc xác định tiếng ồn bổ sung trong giọng nói đang chạy là thách thức lớn hơn đối với hệ thống cơ sở, trong khi mô hình CNN có thể phát hiện hiệu quả sự hiện diện của nền tiếng ồn trong lời nói đang chạy. Cho rằng đối với mỗi tín hiệu nhiễu trong tập dữ liệu, chúng tôi đã chọn một phân đoạn ngẫu nhiên của tệp tiếng ồn và giá trị SNR ngẫu nhiên và âm thanh đó đặc điểm của các tập tin tiếng ồn được sử dụng trong các thí nghiệm này thay đổi theo thời gian (ngoại trừ tiếng ồn trắng), xác suất quan sát các tín hiệu nhiễu bị suy giảm hoàn toàn giống nhau đoạn nhiễu có giá trị SNR tương tự là rất thấp. Vì vậy, dựa trên kết quả, chúng tôi mong đợi hệ thống được đề xuất



Hình 3 Trực quan hóa bản đồ độ mặn, được vẽ bằng màu xanh lam trên phía trên biểu đồ phổ tín hiệu thang màu xám, bằng cách áp dụng điểm-CAM phương pháp phổ log-mel của tín hiệu giọng nói bệnh lý (nguyên âm duy trì /a/) bị biến chất bởi ba kiểu biến chất. Các cường độ của bản đồ độ mặn, được thể hiện trên thang điểm từ 0 đến 255 theo thanh màu, minh họa tầm quan trọng của từng vùng trong không gian đầu vào đến lớp mục tiêu



Hình 4 Trực quan hóa bản đồ độ mặn, được vẽ bằng màu xanh lam trên phía trên biểu đồ phổ tín hiệu thang màu xám, bằng cách áp dụng điểm-CAM phương pháp phổ log-mel của một bài phát biểu thông thường tín hiệu bị suy giảm bởi ba loại suy thoái. Cường độ của bản đồ độ nổi, được hiển thị trên thang điểm từ 0 đến 255 bằng các thanh màu, minh họa tầm quan trọng của từng vùng của không gian đầu vào đối với lớp mục tiêu

để có thể khái quát hóa các loại tiếng ồn không được nhìn thấy trong quá trình giai đoạn đào tạo.

Vì các mô hình học sâu, chẳng hạn như CNN, là những cỗ máy kết hợp có xu hướng học con đường dễ dàng nhất để đạt được điều đó. liên kết dữ liệu đầu vào vào các nhãn, người ta có thể nghi ngờ rằng hiệu suất tốt hơn của phương pháp đề xuất so với hệ thống cơ sở có thể là do nhận được những ảnh hưởng giả từ một số yếu tố gây nhiễu trong dữ liệu. Vì vậy, điều quan trọng là phải hiểu được cơ sở trên đó các mô hình CNN đưa ra quyết định cụ thể về sự hiện diện của sự suy giảm trong một tín hiệu. Có rất nhiều loại các kỹ thuật để hiểu hành vi của các tổ chức phức tạp các mô hình học sâu và cách chúng tạo ra một mô hình cụ thể quyết định [35]. Ánh xạ kích hoạt lớp theo điểm (CAM) là một trong những kỹ thuật ánh xạ biểu diễn nội bộ trong CNN và cung cấp một cách có ý nghĩa, giải thích trực quan chi tiết về phức tạp dựa trên CNN mô hình [20].

Trong phương pháp này, các mặt nạ khác nhau, được gọi là bản đồ kích hoạt, được áp dụng cho hình ảnh đầu vào. phổ log-mel của tín hiệu giọng nói trong các thí nghiệm của chúng tôi. Sau đó, điểm dự đoán cho mỗi lần kích hoạt bản đồ được tính toán và sử dụng như một chỉ báo về tầm quan trọng của bản đồ kích hoạt đó. Bằng cách phủ lên trọng số bản đồ kích hoạt trên hình ảnh đầu vào, các phần của hình ảnh có ảnh hưởng lớn nhất đến điểm số của lớp mục tiêu trong dự đoán của mô hình CNN được làm nổi bật. Hình 3 hiển thị các bản đồ độ nổi được tạo bằng cách áp dụng phương pháp điểm-CAM cho tín hiệu giọng nói bệnh lý bị suy giảm. Chúng ta có thể quan sát sự khác biệt giữa những điểm được đánh dấu các vùng trong ảnh tùy thuộc vào loại suy thoái. Ví dụ, trong Hình 3a, nguyên âm duy trì /a/ là xuống cấp bởi tiếng ồn nhà hàng. Có thể quan sát thấy rằng máy dò tiếng ồn có xu hướng tập trung vào các khu vực có phạm vi rộng trong biểu đồ phổ log-mel trên toàn bộ tần số, cụ thể là trên cả hai vùng chắp vá trong quang phổ log-mel, tương ứng với âm thanh lạch cạch của bộ đồ ăn đĩa và một số vùng tần số thấp tương ứng với tiếng ồn ào trong nhà hàng. Mặt khác tay, như trong Hình 3b và c, các máy dò âm vang và biến dạng có xu hướng tập trung nhiều hơn vào liên tục vùng dọc theo trục thời gian trong biểu đồ phổ log-mel và chủ yếu ở vùng tần số cao. Kết quả cho thấy các vùng tần số cao có ảnh hưởng nhiều hơn và quan trọng trong việc xác định độ méo và âm vang. Tuy nhiên xu hướng lại hoàn toàn trái ngược trong nhận dạng người nói, điều này có tầm quan trọng lớn đối với vùng tần số thấp (khoảng 200 Hz đến 3 kHz) [36].

Các bản đồ độ nổi được tạo ra bằng cách áp dụng phương pháp Score-CAM cho tín hiệu giọng nói đang chạy bình thường bị suy giảm được hiển thị trong Hình 4. Chúng ta có thể quan sát thấy xu hướng, tức là bộ phát hiện tiếng ồn tập trung vào các khu vực trên toàn bộ tần số và các tần số khác tập trung vào vùng tần số cao, giống như giọng nói bệnh lý. Điều thú vị là có thể thấy máy dò méo chủ yếu tập trung vào vùng tần số cao của các khung được lồng tiếng (vùng công suất cao) trong chương trình đặc tả log-mel. Đó là lý do tại sao người ta cho rằng phi tuyến sự biến dạng xuất hiện đáng chú ý khi giọng nói gốc trở nên ồn ào.

Bảng 4 Tác động của việc thay đổi tần số cắt thấp hơn và cao hơn của biểu đồ phổ log-mel đối với hiệu suất của từng loại máy dò tín hiệu giọng nói bệnh lý

Máy dò	dòng chảy [Hz]				bay [kHz]			
	0	300	700	2500	4.3	11	15	người đi theo chủ nghĩa Nyquist
Tiếng ồn	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00	1,00±0,00
Méo mó	0,99±0,00	0,98±0,01	0,98±0,01	0,98±0,01	0,97±0,01	0,98±0,01	0,97±0,01	0,99±0,00
Tiếng vang	0,93±0,01	0,93±0,01	0,91±0,01	0,86±0,02	0,83±0,02	0,88±0,01	0,92±0,01	0,93±0,01

Các kết quả ở dạng trung bình khoảng tin cậy AUC ±95%

Bảng 5 Tác động của việc thay đổi tần số cắt thấp hơn và cao hơn của biểu đồ phổ log-mel đối với hiệu suất của từng loại máy dò tín hiệu giọng nói đang chạy bình thường

Máy dò	dòng chảy [Hz]				bay [kHz]			
	0	300	700	2500	4.3	11	15	ngữ đi theo chủ nghĩa Nyquist
Tiếng ồn	0,95±0,01	0,92±0,01	0,94±0,00	0,91±0,01	0,99±0,00	1,00±0,00	0,99±0,00	0,95±0,01
Méo mó	1,00±0,00	1,00±0,00	0,99±0,00	0,99±0,00	0,99±0,01	1,00±0,00	0,99±0,00	1,00±0,00
Tiếng vang	0,99±0,00	0,99±0,00	0,99±0,00	0,99±0,00	0,90±0,01	0,90±0,01	0,97±0,01	0,99±0,00

Các kết quả ở dạng trung bình khoảng tin cậy AUC ±95%

Để nghiên cứu sâu hơn về tầm quan trọng của tần số cao khu vực trong việc xác định suy thoái, chúng tôi đã đánh giá hiệu suất của phương pháp được đề xuất bằng cách sử dụng log-mel đồ thị có tần số cắt khác nhau. log-mel quang phổ thường được bắt nguồn bằng cách áp dụng hình tam giác các bộ lọc được căn chỉnh theo các khoảng đều nhau theo thang đo mel với công suất STFT chuẩn hóa. Tần số tuyến tính f tính bằng Hz có thể được chuyển đổi sang tần số thang đo mel m bằng cách sử dụng phương trình sau [37]:

$$m = (f) = \ln 1 \frac{1000}{+ \ln(1 + 1000/700)} \quad \frac{f}{700} \cdot \quad (7)$$

Chúng tôi xác định tần số cắt thấp và cao cho bộ lọc thang đo mel như mlow = (flow) và mhigh = (fhigh), tương ứng. Hiệu suất của từng máy dò đối với giọng nói bệnh lý và giọng nói bình thường khi thay đổi giá trị của tần số cắt được báo cáo trong Bảng 4 và 5, tương ứng. Trong các bảng này, tần số được hiển thị ở thang đo Hz tuyến tính có thể được chuyển đổi sang thang đo mel sử dụng phương trình. (7). Cần lưu ý rằng, theo nghĩa của phương trình. (7), lượng tần số mel có trong các dải tần nhỏ hơn 300 Hz, 700 Hz và 2,5 kHz tương ứng với các giá trị nằm trong dải tần tương ứng hơn 15 kHz, 11 kHz và 4,3 kHz. Bất chấp thực tế này, hiệu suất của máy dò âm vang trở nên kém hơn đáng kể do giảm tần số cao tới 11 kHz. Tuy nhiên, việc tăng lưu lượng chỉ có tác động hạn chế đến hiệu suất của máy dò âm vang. Ngược lại, hiệu suất của tiếng ồn máy dò trở nên tốt hơn một chút bằng cách giảm f, có lẽ là do sự gia tăng độ phân giải thấp hơn các vùng tần số. Trong khi đó, hiệu suất của máy dò biến dạng vẫn gần như giữ nguyên ngay cả khi thay đổi tần số cắt cao hơn và thấp hơn. Những kết quả này rất phù hợp với những giải thích trực quan trong các thí nghiệm trước đó và chỉ ra rằng 8 kHz điển hình của tốc độ lấy mẫu lấy từ hệ thống điện thoại không đủ để xác định tiếng vang. Chúng tôi suy luận rằng mức cao tần số âm thanh có xu hướng dễ dàng bị suy giảm bởi một bức tường hoặc các vật cản trở khác trong phòng và kết quả là giảm chấn xuất hiện ở vùng tần số cao.

5. Kết luận

Trong bài báo này, chúng tôi đã đề xuất một phương pháp mới dựa trên CNN Phương pháp xác định sự suy giảm tín hiệu tiếng nói Trong phương pháp này, một tập hợp các mô hình CNN, mỗi mô hình chịu trách nhiệm phát hiện một loại suy thoái cụ thể, có đã được sử dụng. Ưu điểm của phương pháp này so với các phương pháp phát hiện sự suy giảm đa lớp là tính năng song song và Các bộ dò độc lập tạo điều kiện thuận lợi cho cả việc phát hiện sự hiện diện của sự kết hợp các suy giảm trong tín hiệu giọng nói. và loại bỏ một ngoại lệ của một loại suy thoái mới cho mà các mô hình chưa được đào tạo. Các CNN đã được huấn luyện bằng phổ log-mel của một số lượng lớn tín hiệu giọng nói bị suy giảm. Kết quả thực nghiệm sử dụng hai loại giọng nói khác nhau, đó là nguyên âm duy trì bệnh lý và giọng nói chạy bình thường Hiệu quả của phương pháp đề xuất trong việc phát hiện sự suy giảm tín hiệu tốt hơn hệ thống hiện đại. Hơn nữa, bằng cách sử dụng kỹ thuật Score-CAM, chúng tôi đã giải thích một cách trực quan cách các mô hình CNN tạo ra một quyết định cụ thể trong việc xác định sự suy giảm tín hiệu. Nó cũng tiết lộ rằng các vùng tần số cao trong log-mel quang phổ mang thông tin quan trọng để xác định tiếng vang. Nó làm cho việc xác định tiếng vang thách thức khi áp dụng tín hiệu chất lượng điện thoại của Tần số lấy mẫu 8 kHz.

Các từ viết tắt

CNN: Mạng lưới thần kinh tích chập; RIR: Đáp ứng xung của phòng; GMM: mô hình hỗn hợp Gaussian; CAM: Ánh xạ kích hoạt lớp; BN: Lô bình thường hóa; ReLU: Đơn vị tuyến tính chỉnh lưu; STFT: Biến đổi phạm vi thời gian ngắn; SGD: Độ dốc giảm dần ngẫu nhiên; UBM: Mô hình nền phổ quát; MFCC: Hệ số epstral tần số Mel; MMPD: mPower di động Parkinson bệnh; PD: Bệnh Parkinson; SNR: Tỷ lệ tín hiệu trên tạp âm; dB: Đêxiben; Hz: Hertz; CELP: Dự đoán tuyến tính kích thích bằng mã; NS: Ồn ào; NRS: Tiếng ồn vang dội; ROC: Đặc tính hoạt động của máy thu; AUC: Diện tích dưới đường cong; CV: Chữ Thập Thẩm định.

Sự nhìn nhận

Không áp dụng được.

Tác giả đóng góp

YS đã tiến hành các thí nghiệm. AHP và MGC đã thiết kế ý tưởng cho bản thảo. Tất cả các tác giả đã đóng góp vào việc viết tác phẩm này. Hơn thế nữa, tất cả các tác giả đã đọc và phê duyệt bản thảo cuối cùng.

Kinh phí

Công trình này được tài trợ bởi Quỹ nghiên cứu độc lập Đan Mạch: DFF 4184-00056.

Tính sẵn có của dữ liệu và tài liệu Bộ
dữ liệu đư ợc sử dụng để xác định sự suy giảm giọng nói bệnh lý có sẵn trên Cổng
thông tin nghiên cứu công cộng mPower https://www.synapse.org/#!
Khớp thần kinh:syn4993293/. Bộ dữ liệu đư ợc sử dụng để xác định sự suy giảm chất
lư ợng trong giọng nói thông thư ờng hiện có tại Trung tâm Nghiên cứu Công nghệ
Giọng nói tại Đại học Edinburgh https://datashare.is.ed.ac.uk/handle/ 10283/2791
và https://dx.doi.org/10.7488/ds/2139. Cơ sở dữ liệu về phản hồi xung của phòng có
tại đây http://www.ind.rwth-aachen.de/air. Danh sách dữ liệu đư ợc lấy mẫu có sẵn từ
tác giả tư ơ ng ứng theo yêu cầu hợp lý.

Cạnh tranh lợi ích Các
tác giả tuyên bố rằng họ không có lợi ích cạnh tranh.

Đã nhận: ngày 12 tháng 10 năm 2020 Đư ợc chấp nhận: ngày 15 tháng 1 năm 2021
Published online: 05 February 2021

Ngư ời giới thiệu

1. S. Ghai, R. Sinha, Cắt bớt tính năng thích ứng để giải quyết vấn đề âm thanh không khớp trong việc tự động nhận dạng giọng nói của trẻ em. APSIPA Trans. Thông tin tín hiệu Quá trình. 5, 1–13 (2016)
2. A. Alexander, F. Botti, D. Dessimoz, A. Drygajlo, Ảnh hư ởng của các điều kiện ghi không khớp đối với khả năng nhận dạng con ngư ời và ngư ời nói tự động trong các ứng dụng pháp y. Khoa học pháp y. Int. 146, 95–99 (2004)
3. V. Mitra, A. Tsiartas, E. Shriberg, trong Hội nghị Quốc tế về Âm học, Xử lý giọng nói và tín hiệu (ICASSP). Hiệu ứng tiếng ồn và âm vang trong việc phát hiện trầm cảm từ lời nói, (2016), trang 5795–5799 4. AH Poorjam, MS Kavalekalam, L. Shi, JP Raykov, JR Jensen, MA Little, MG Christensen, Kiểm soát chất lư ợng tự động và cải tiến để phát hiện bệnh Parkinson từ xa dựa trên giọng nói. Bài phát biểu của Cộng đồng. 127, 1–16 (2021)
5. M. Fakhry, AH Poorjam, MG Christensen, tại Hội nghị xử lý tín hiệu châu Âu (EUSIPCO). Cải thiện giọng nói bằng cách phân loại các tín hiệu nhiễu đư ợc phân tách bằng bộ lọc NMF và Wiener, (2018), trang 16–20 JHL Hansen, A. Kumar, P.
6. Angkititrakul, Bù không khớp môi trư ờng bằng các phư ơ ng pháp dựa trên không gian riêng trung bình để nhận dạng giọng nói mạnh mẽ. Int. J. Công nghệ giọng nói. 17(4), 353–364 (2014)
7. BW Gillespie, HS Malvar, DAF Florencio, tại Hội nghị quốc tế của IEEE về Âm học, Xử lý giọng nói và tín hiệu (ICASSP). Giảm âm vang giọng nói thông qua lọc thích ứng băng con kurtosis tối đa, tập. 6, (2001), trang 3701–3704 H. Kun, W. Yuxuan, W.
8. DeLiang, SW William, M. Ivo, Z. Tao, Học lập bản đồ quang phổ để khử tiếng nói và khử nhiễu. IEEE Trans. Âm thanh lời nói Lang. Quá trình. 23(6), 982–992 (2015)
9. JS Abel, trong Hội nghị quốc tế của IEEE về Âm học, Lời nói và Xử lý Tín hiệu. Khôi phục tín hiệu bị cắt IEEE Computer Society, (1991), trang 1745–1748 10. B. Defraene, N. Mansour, S. De Hertogh, T. Van Waterschoot, M. Diehl, M. Moonen, Tách tín hiệu âm thanh bằng cách sử dụng cảm biến nén theo cảm nhận. IEEE Trans. Âm thanh lời nói Lang. Quá trình. 21(12), 2627–2637 (2013)
11. DS Williamson, D. Wang, Che giấu tần số thời gian trong miền phức tạp để khử tiếng nói và khử nhiễu. Chuyển đổi IEEE/ACM. Âm thanh lời nói Lang. Quá trình. 25(7), 1492–1501 (2017)
12. T. Dietzen, S. Doclo, M. Moonen, T. van Waterschoot, Loại bỏ búp sóng bên tích hợp và dự đoán tuyến tính Bộ lọc Kalman để khử tiếng nói chung bằng nhiều micrô, khử tiếng nói gây nhiễu và giảm tiếng ồn. Chuyển đổi IEEE/ACM. Âm thanh lời nói Lang. Quá trình. 28, 740–754 (2020)
13. I. Kodrasi, S. Doclo, Giảm âm vang chung và giảm tiếng ồn dựa trên cân bằng âm thanh đa kênh. Chuyển đổi IEEE/ACM. Âm thanh lời nói Lang. Quá trình. 24(4), 680–693 (2016)
14. L. Ma, DJ Smith, BP Milner, tại Hội nghị diễn thuyết châu Âu lần thứ 8 Truyền thông và Công nghệ. Nhận thức bối cảnh sử dụng phân loại tiếng ồn môi trư ờng, (2003), trang 1–4 15. JM Desmond, LM Collins, CS Throckmorton, Sử dụng mô hình thống kê theo kênh cụ thể để phát hiện âm vang trong kích thích cấy ốc tai điện tử. J. Âm thanh. Sóc. Là. 134(2), 1112–1120 (2013)
16. SV Aleinik, MY Nikolaevich, SA Vladimirovich, Phát hiện các đoạn bị cắt trong tín hiệu âm thanh. J. Khoa học. Công nghệ. Thông tin Technol. Máy móc. Opt. 92(4), 91–97 (2014)
17. F. Bie, D. Wang, J. Wang, TF Zheng, Phát hiện và tái tạo lại lời nói bị cắt để nhận dạng ngư ời nói. Bài phát biểu của Cộng đồng. 72, 218–231 (2015)
18. AH Poorjam, JR Jensen, MA Little, MG Christensen, trong Kỷ yếu Hội nghị thư ờng niên của Hiệp hội Truyền thông Lời nói Quốc tế, InterSpeech. Phân loại biến dạng chiếm ư u thế để xử lý trư ớc các nguyên âm trong phân tích giọng nói y sinh từ xa, (2017), trang 289–293 19. AH Poorjam, MA Little, JR Jensen, MG Christensen, năm 2018 Hội nghị quốc tế của IEEE về Âm học, Lời nói và Xử lý Tín hiệu (ICASSP). Một cách tiếp cận tham số để phân loại các biến dạng trong giọng bệnh lý, (2018), trang 286–290 20. H. Wang, M. Du, F. Yang, Z. Zhang, Score-CAM: hình ảnh đư ợc cải thiện giải thích thông qua ánh xạ kích hoạt lớp theo điểm số. Hội nghị IEEE/CVF 2020 về Hội thảo Nhận dạng Mẫu và Thị giác Máy tính (CVPRW), 111–119 (2020)
21. K. Simonyan, A. Zisserman, trong Hội nghị Quốc tế lần thứ 3 về Trình bày Học tập, ICLR 2015 - Kỷ yếu Theo dõi Hội nghị. Mạng tích chập rất sâu để nhận dạng hình ảnh quy mô lớn, (2015), trang 1–14
22. J. Bjorck, C. Gomes, B. Selman, KQ Weinberger, trong Những tiến bộ trong hệ thống xử lý thông tin thần kinh (NeurIPS). Tìm hiểu về chuẩn hóa hàng loạt, (2018), trang 7694–7705
23. BC Moore, Giới thiệu về Tâm lý Thính giác. (Ngọc lục bảo, Bingley, 2012)
24. K. Choi, G. Fazekas, M. Sandler, K. Cho, tại Hội nghị xử lý tín hiệu châu Âu lần thứ 26 năm 2018 (EUSIPCO). So sánh các phư ơ ng pháp tiền xử lý tín hiệu âm thanh cho mạng lư ới thần kinh sâu về gắn thẻ âm nhạc, (Rome, 2018), trang 1870–1874 25. BM Bot, C. Suver, EC Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, ER Dorsey, SH Friend, AD Trister, Nghiên cứu mPower, dữ liệu di động về bệnh Parkinson đư ợc thu thập bằng ResearchKit. Dữ liệu khoa học. 3, 1–9 (2016)
26. AK Ho, R. Iansek, C. Marigliani, JL Bradshaw, S. Gates, Bài phát biểu suy giảm ở một mẫu lớn bệnh nhân mắc bệnh Parkinson. Cư xử. Thần kinh. 11(3), 131–137 (1998)
27. IR Titze, DW Martin, Nguyên tắc sản xuất giọng nói. Âm thanh. Sóc. Là. 104(3), 1148 (1998)
28. C. Valentini-Botinhao, Cơ sở dữ liệu tiếng ồn để luyện giọng nói các thuật toán nâng cao và mô hình tts [trực tuyến] (2017). Có sẵn: http://dx.doi.org/10.7488/ds/2117 29. M. Jeub, M. Schäfer, H. Krüger, C. Nelke, C. Beaugeant, P. Vary, ở Proc. Int. Đại hội Âm học (ICA), Sydney, Australia. Chúng ta có cần giảm âm vang cho điện thoại cầm tay không? (2010), trang 1–7
30. MR Schroeder, Phư ơ ng pháp đo thời gian vang mới. J. Âm thanh. Sóc. Là. 37(6), 1187–1188 (1965)
31. MR Schroeder, BS Atal, tại Hội nghị quốc tế của IEEE về Âm học, Xử lý giọng nói và tín hiệu (ICASSP). Dự đoán tuyến tính kích thích bằng mã (CELP): giọng nói chất lư ợng cao ở tốc độ bit rất thấp, vol. 10, (1985), trang 937–940 32. C. Valentini-Botinhao, Cơ sở dữ liệu tiếng nói vang dội ồn ào để đào tạo các thuật toán nâng cao giọng nói và mô hình tts [trực tuyến] (2017). Có sẵn: https://dx.doi.org/10.7488/ds/2139 33. J. Thiemann, N. Ito, E. Vincent, Môi trư ờng đa dạng đa kênh Cơ sở dữ liệu tiếng ồn âm thanh: cơ sở dữ liệu về các bản ghi tiếng ồn môi trư ờng đa kênh. J. Âm thanh. Sóc. Là. 133(5), 3591–3591 (2013)
34. SL Smith, PJ Kindermans, C. Ying, QV Le, ở Lớp 6 Quốc tế Hội nghị về đại diện học tập. Đừng giảm tốc độ học, hãy tăng quy mô lô, (2018), trang 1–11 35. M. Du, N. Liu, X. Hu, Kỹ thuật học máy có thể diễn giải. Cộng đồng. ACM. 63(1), 68–77 (2019)
36. X. Lu, J. Đặng, Nghiên cứu sự phụ thuộc giữa tần số các thành phần và đặc điểm của loa để nhận dạng loa độc lập với văn bản. Bài phát biểu của Cộng đồng. 50(4), 312–322 (2008)
37. J. Harrington, S. Cassidy, Kỹ thuật âm học lời nói, tập. 8. (Springer Science & Business Media, Hà Lan, 2012)

Ghi chú của nhà xuất bản