Phát hiện giọng nói với mạng kết hợp CNN-BiLSTM

Nguyễn Đăng Hải*‡, Nguyễn Gia Huy*‡, Phạm Minh Hoàng†‡, Đỗ Đức Hào†‡
*Trường Đại Học Bách Khoa - Đại Học Quốc Gia TPHCM
†Trường Đại Học Khoa Học Tự Nhiên - Đại Học Quốc Gia TPHCM
‡ Công ty CP Công nghệ OLLI
Email: hai.nguyen.dang.1307@hcmut.edu.vn, {huy.nguyen, hoang.pham, hao}@olli-ai.com
Liên hệ: hao@olli-ai.com

***Tóm tắt nội dung*—Phát hiện giọng nói là một bước tiền xử lý quan trọng trong những mô hình trợ lý ảo. Bài toán cải thiện độ chính xác mô hình phát hiện giọng nói là một chủ đề đã có từ rất lâu. Sự phát triển của các mô hình mạng lưới thần kinh nhân tạo đã đóng góp đáng kể cho những mô hình phát hiện giọng nói. Mạng Convolution Neuron Network (CNN) đã được áp dụng nhiều và cho ra kết quả khả quan. Đã có nhiều đề xuất sử dụng mô hình kết hợp với mạng CNN dùng để cải tiến hệ thống phát hiện giọng nói, và trong bài báo này, chúng tôi sử dụng mạng CNN kết hợp với mạng trí nhớ dài-ngắn hạn 2 chiều (BiLSTM) dùng để cải thiện mô hình phát hiện giọng nói tiếng Việt. Mô hình sử dụng bộ dữ liệu VIVOS để huấn luyện và cho kết quả khả quan với độ cải thiện của mô hình khoảng 9.5% khi đánh giá trên tập AVA-Speech.**

***Keywords*-mạng thần kinh nhân tạo; học sâu; phát hiện giọng nói; Convolution Neural Network; Bidirectional long short-term memory.**

I. Giới thiệu

Phát hiện giọng nói (Voice Activity Detection - VAD) được tạo ra với mục đích phân loại những đoạn âm thanh có giọng nói và những đoạn âm thanh không có giọng nói. VAD đóng một vai trò quan trọng, nó là bước tiền xử lý cho nhiều hệ thống xử lý giọng nói, cải thiện giọng nói (speech enhancement), nhận dạng người nói và nhận dạng tiếng nói tự động (ASR). Bằng cách xem bài toán VAD như là bài toán phân loại trên từng khung hình, nhiều bộ phân lớp có thể được huấn luyện để nhận dạng những khung hình nào có chứa tiếng nói hay không chứa tiếng nói.

Các kiến trúc support vector machine (SVM) đã được sử dụng rộng rãi [1]–[4], và gần đây là các kiến trúc mạng neural [5]–[10]. Mạng Convolution Neural Network (CNN) là một kiến trúc mạng nơ-ron rất phổ biến trong việc nhận dạng và phân loại, và đã được sử dụng rất nhiều trong bài toán VAD [11]–[13]. Tuy nhiên, việc thiếu mô hình tính tuần tự dữ liệu theo dạng chuỗi gây ra hiện tượng giảm hoặc chèn khung hình trong các đoạn có hoặc không có lời nói [14]. Nhằm giảm thiểu vấn đề trên, [14] cũng đã đề xuất đến việc sử dụng mô hình kết hợp giữa mạng nơ-ron tích chập và mạng trí nhớ ngắn hạn định hướng dài hạn (CNN-BiLSTM).

Trong bài báo này, chúng tôi sử dụng bộ dữ liệu tiếng Việt [15] dùng để huấn luyện mô hình CNN-BiLSTM, sau đó so sánh mô hình này với các mô hình CNN và CNN-LSTM trên tập AVA-Speech được trình bày ở phần IV-C. Các đóng góp chính của chúng tôi trong bài báo này bao gồm:

- So sánh chất lượng của mô hình CNN với mô hình BiLSTM.
- Huấn luyện mô hình trên tập dữ liệu tiếng Việt VIVOS và đánh giá mô hình trên 2 tập đánh giá: AVA-Speech và VIVOS, nhằm kiểm tra đặc trưng âm thanh của tiếng Việt và không phải tiếng Việt.

Phần còn lại của bài báo được trình bày theo thứ tự: phần II trình bày về Log-Mel Filterbank Energy, và cách các mô hình CNN và BiLSTM được triển khai trong VAD. Phần III trình bày về mô hình đề xuất. Kết quả thực nghiệm được trình bày ở phần IV và cuối cùng là phần kết luận của bài báo ở phần V.

II. Các nghiên cứu liên quan

A. Log mel-filterbank Energy

Có nhiều phương pháp để biểu diễn âm thanh dưới dạng hình ảnh. Trong bài báo này, dữ liệu đầu vào mà chúng tôi chọn cho mô hình CNN là log mel-filterbank energy đã được đề cập ở [16]. Nghiên cứu tại [17] đã cho thấy mel-scaled short time Fourier transform (STFT) spectrograms biểu diễn tốt hơn linear-scaled STFT spectrograms, constant-Q transform (CQT) spectrogram, continuous Wavelet transform (CWT) scalogram và MFCC

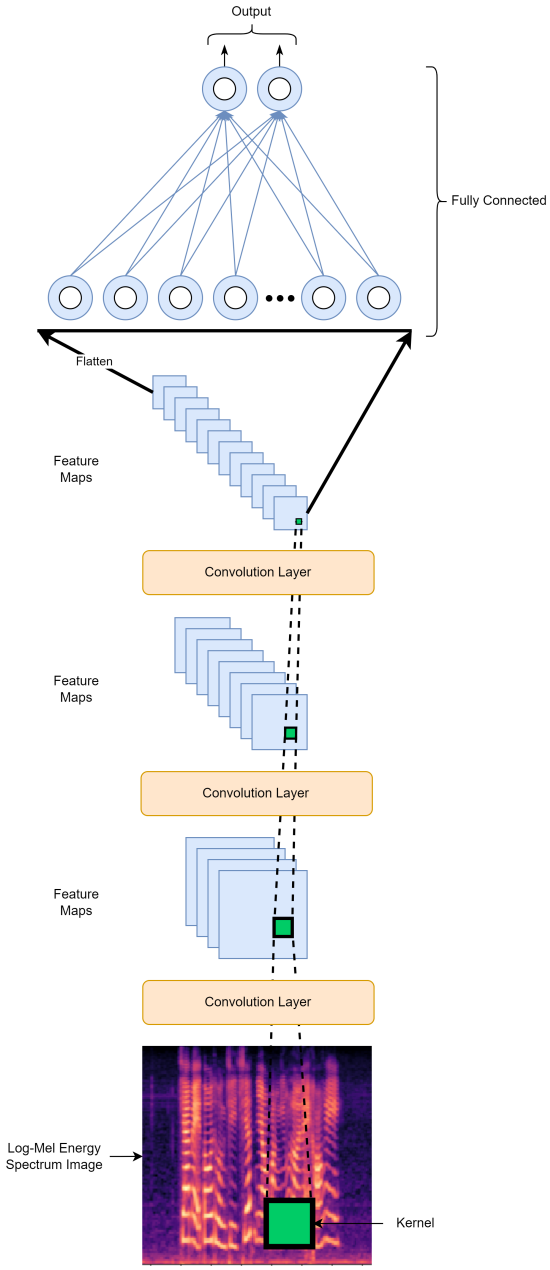
cepstrogram làm dữ liệu đầu vào CNN trong bài toán VAD. Thêm vào đó, [16] cho thấy việc sử dụng log mel-filterbank energy được trích xuất từ mel-scaled STFT spectrogram biểu diễn tốt hơn khi sử dụng cho CNN so với những cái còn lại. Thêm vào đó, log mel-filterbank energy feature có ít hệ số hơn trên mỗi khung hình (frame) so với linear-scaled STFT spectrogram và mel-scaled STFT spectrogram, giúp giảm thời gian thực thi và thu gọn kiến trúc mô hình CNN.

B. Convolution Neural Network

CNN được sử dụng cho việc nhận dạng và phân loại. Mô hình mạng CNN được biểu diễn ở hình 1. Mô hình nhận Log mel energy Spectrum Images làm dữ liệu đầu vào, kèm theo là các lớp tích chập và các lớp tổng hợp với một lớp kết nối đầy đủ. Các lớp tích chập có khả năng trích xuất thông tin từ dữ liệu đầu vào nhờ vào các kernel học tập có trọng số và các hàm phi tuyến. Những kernel này được nhân tích chập trên toàn bộ dữ liệu đầu vào, và sinh ra các feature map. Thông tin của feature map có thể được đơn giản hóa bằng cách sử dụng max-pooling nhằm giảm độ tính toán, và tiếp tục đẩy nó vào lớp tích chập kế tiếp. Sau đó, lớp kết nối đầy đủ (fully-connected layer) sẽ duỗi tất cả các kênh được sinh ra từ lớp tích chập cuối cùng nhằm phân loại bằng cách sử dụng một lớp đầu ra phi tuyến. Cuối cùng, hàm kích hoạt softmax được sử dụng để tính toán xác suất đại diện cho những đoạn có hay không có tiếng nói.

C. Bidirectional LSTM

Mô hình Recurent Neural Network (RNN) được sử dụng trong những bài toán có dữ liệu đầu vào tuần tự. Tuy nhiên, RNN không thể ghi nhớ thông tin dài hạn, do đó, mạng LSTM ra đời nhằm giải quyết nhược điểm này. Kiến trúc của LSTM dựa trên việc sử dụng những ô nhớ để ghi nhớ thông tin dài hạn và điều chỉnh thông qua cơ chế cổng. Một mô hình LSTM thông thường sẽ có 3 loại cổng: cổng vào i_t, cổng quên f_t và cổng ra o_t. Ba cổng này được biểu diễn ở hình 2, sử dụng phép tính pointwise và hàm sigmoid để kiểm soát trạng thái của ô nhớ. Dữ liệu vào x_t (trạng thái hiện tại) và dữ liệu ra h_{t-1} từ trạng thái ẩn của lớp trước đó được vào tất cả các cổng. Cổng quên quyết định những thông tin nào nên loại bỏ hoặc giữ lại. Hàm sigmoid chuyển trạng thái của i_t và h_{t-1} thành giá trị nằm trong khoảng 0 và 1. Nếu giá trị này càng lớn thì lượng thông tin từ quá khứ sẽ được giữ lại

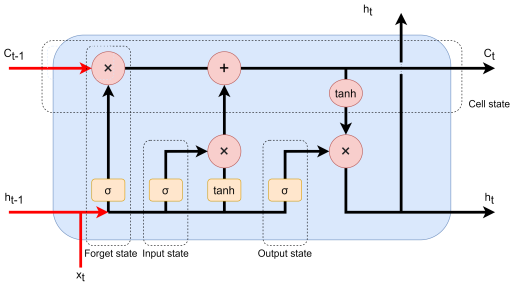


Hình 1: Mô hình CNN trong VAD

càng nhiều, và ngược lại. Công thức tính giá trị cổng quên như sau, trong đó σ là hàm sigmoid, W và b lần lượt là trọng số và bias của cổng:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

Sau đó, cổng vào sẽ nhận x_t và h_{t-1} vào hàm



Hình 2: Sơ đồ mô hình BiLSTM

sigmoid và hoạt động tương tự như cổng quên. Bằng cách biến đổi thành giá trị từ 0 đến 1, cổng vào quyết định bao nhiêu lượng thông tin đầu vào sẽ ảnh hưởng đến trạng thái mới. Công thức tính của cổng vào như sau:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

Tiếp theo, x_t và h_{t-1} sẽ được truyền vào hàm $tanh$ để cập nhật cho ô nhớ trạng thái $\hat{C}_t$ và C_t lần lượt theo công thức 3 và 4.

$$\hat{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{3}$$

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t \tag{4}$$

Mô hình sẽ tiếp tục tính toán trạng thái ẩn cho bước tiếp theo bằng công thức 5 và 6.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = \tanh(C_t) * o_t \tag{6}$$

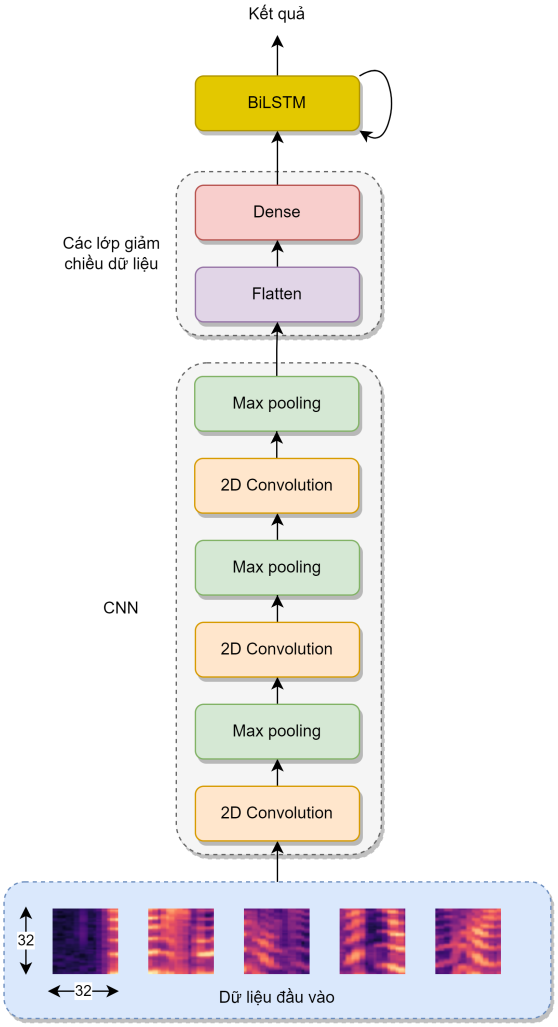
LSTM thông thường chỉ xử lý thông tin theo một chiều (thông tin đến trước). Trong khi đó, kiến trúc BiLSTM có 2 lớp LSTM lần lượt sử dụng ngữ cảnh bên trái và bên phải. Hình 2 mô tả kiến trúc của BiLSTM. LSTM đầu nhận chuỗi dữ liệu đầu vào từ quá khứ và LSTM còn lại nhận chuỗi dữ liệu đầu vào từ tương lai, và sau đó kết hợp cả 2 vector trạng thái ẩn $\overleftarrow{h_t}$ và $\overrightarrow{h_t}$ lại với nhau.

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t} \tag{7}$$

BiLSTM cho kết quả tốt hơn LSTM vì nó sử dụng thông tin của cả trước và sau.

III. Mô hình đề xuất

Mô hình kết hợp CNN-BiLSTM được chúng tôi sử dụng cho VAD được mô tả trong hình 4. Nó bao gồm ba lớp CNN 2D, với hàm kích hoạt ReLU và ba lớp max-pooling. Để giảm chiều dữ liệu, kết quả đầu ra từ lớp max-pooling thứ 3 được làm phẳng (flatten) và được đưa vào lớp kết nối đầy đủ với hàm kích hoạt ReLU. Sau đó nó được kết nối tới một lớp BiLSTM với các hàm kích hoạt tanh và các hàm hồi quy sigmoid. Cuối cùng, BiLSTM được đưa qua lớp softmax để tính phân phối xác suất cho các nhãn có hoặc không có tiếng nói.



Hình 3: Mô hình CNN-BiLSTM

Mô hình nhận chuỗi ảnh spectrogram kích thước 32×32 làm dữ liệu đầu vào, tương tự [9]. Những spectrogram được tạo bằng cách tính log mel-

filterbank energy 32 chiều sử dụng bước nhảy khung hình là 10 ms, và xếp chồng chúng với nhau trên 320 ms để tạo thành một hình ảnh đầu vào, hình ảnh minh hoạ ở hình 4. Lí do chúng tôi chọn log mel-filterbank energy đã được trình bày ở phần II-A.

IV. Thực nghiệm

A. Bộ dữ liệu

Bộ dữ liệu được dùng trong thực nghiệm với dữ liệu có tiếng nói là VIVOS được mô tả ở [15]. Bộ dữ liệu chúng tôi sử dụng bao gồm 45 phút dữ liệu lời nói được thu âm trong môi trường yên tĩnh, và 45 phút dữ liệu lời nói trong môi trường có tạp âm với SNR lần lượt là -5, 0 , 5 và 10 dB. Bộ dữ liệu có lời nói của 19 người (12 nam, 7 nữ). Tần số lấy mẫu là 16 kHz. Với dữ liệu không có tiếng nói, chúng tôi thu được 20 phút được lấy từ freesound.org, bao gồm 40 đoạn âm thanh không có tiếng nói dài 30 giây. Cuối cùng chúng tôi sinh ngẫu nhiên âm thanh từ dữ liệu ở trên và thu được hơn 14 giờ âm thanh để huấn luyện cho mô hình. Tập đánh giá chúng tôi sử dụng là tập AVA-Speech được cắt ra 90 giây cho mỗi 95 đoạn phim, tổng cộng hơn 2,4 giờ. Ngoài ra, chúng tôi cũng đánh giá mô hình trên 5 giờ âm thanh được sinh ngẫu nhiên từ bộ dữ liệu trên làm tập đánh giá mô hình cho tiếng Việt, kết quả thu được lần lượt ở bảng III và IV.

B. Các tham số của mô hình

Các tham số của mô hình được trình bày ở bảng I. Mô hình được huấn luyện qua 25 epoch với kích thước mỗi batch size là 64. Thuật toán tối ưu được sử dụng là thuật toán $adam$ với $\alpha = 0.001, (\beta_1, \beta_2) = (0.9, 0.999), \epsilon = 1e-07$.

Loại	Kích cỡ	Mô tả
	32x32x1	Dữ liệu đầu vào
Convolution	28x28x64	5x5 convolution, 64 filter
Pooling	14x14x64	2x2 max-pooling
Convolution	12x12x128	3x3 convolution, 128 filter
Pooling	6x6x128	2x2 max-pooling
Convolution	4x4x128	3x3 convolution, 128 filter
Pooling	2x2x128	2x2 max-pooling
Flatten	512	Flatten
Fully-connected	128	Fully-connected
BiLSTM	256	BiLSTM
Softmax	2	Softmax

Bảng I: Kiến trúc được dùng để huấn luyện mô hình

C. Kết quả

Chúng tôi sử dụng tập AVA-speech để đánh giá mô hình. AVA-speech là một bộ dữ liệu mở, gồm nhũng đoạn phim đã được gắn nhãn. Tại thời điểm viết bài, chúng tôi thu được 95 đoạn phim được lưu trên YouTube, mỗi đoạn dài 15 phút, tổng thời gian được đánh nhãn là 23 giờ 45 phút. Những đoạn phim bao gồm 4 loại nhãn: "No Speech", "Clean Speech", "Speech+Music", "Speech+Noise".

Label	Time(%)	Segments(%)	AvgDur(s)
CleanSpeech	15.15	16.78	3.20
Speech+Music	25.30	26.00	3.45
Speech+Noise	13.92	13.34	3.70
NoSpeech	45.63	43.88	3.68

Bảng II: Thông số của bộ dữ liệu AVA-Speech

AVA-Speech cung cấp một bộ âm thanh đa dạng về người nói, điều kiện âm thanh cũng như ngôn ngữ. Tập dữ liệu có thời gian có tiếng nói và không có tiếng nói cũng xấp xỉ bằng nhau, và hầu hết các đoạn âm thanh có tiếng nói đều bị nhiễu.

Model	Clean	Noise	Music	All
CNN	0.523	0.521	0.542	0.527
CNN-LSTM	0.525	0.510	0.522	0.517
CNN-BiLSTM	**0.591**	**0.573**	**0.572**	**0.577**

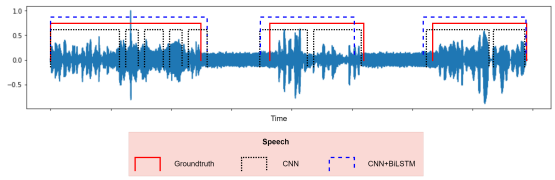
Bảng III: Kết quả TPR tại FPR=0.315 thu được của 3 model đánh giá trên tập AVA-Speech

Model	Clean	Noise	Music	All
CNN	0.877	0.838	0.832	0.845
CNN-LSTM	0.864	0.845	0.851	0.853
CNN-BiLSTM	**0.891**	**0.865**	**0.861**	**0.873**

Bảng IV: Kết quả TPR tại FPR=0.315 thu được của 3 model đánh giá trên tập VIVOS

Kết quả của 3 loại mô hình khi đánh giá với AVA-Speech được biểu diễn ở bảng III, với các điều kiện "Clean Speech", "Speech+Noise", "Speech+Music". Bảng III cho thấy kết quả trung bình của CNN-BiLSTM đều tốt hơn so với mô hình CNN và mô hình CNN-LSTM. Tổng thể mô hình CNN-BiLSTM tốt hơn khoảng 9.5% so với mô hình CNN khi đánh giá trên tập AVA-Speech. Tuy nhiên, tổng quan thì 3 mô hình trên vẫn chưa cho kết quả tốt. Trái lại, với tập dữ liệu VIVOS, cả 3 mô hình trên đều cho kết quả khá khả quan. Nhìn chung, mô hình CNN-BiLSTM cho kết quả tốt hơn so với 2 mô hình còn lại, với độ chênh lệch khoảng 5%. Sự khác nhau giữa kết quả khi đánh giá tập dữ liệu

AVA-Speech và VIVOS có thể thấy rằng có sự khác nhau về đặc trưng âm thanh giữa tiếng Việt so với những ngôn ngữ khác.



Hình 4: Kết quả đầu ra của 2 mô hình CNN và CNN-BiLSTM, trong đó, kết quả của mô hình CNN bị "phân mảnh"

Ngoài ra, chúng tôi phát hiện rằng mô hình CNN có hiện tượng "phân mảnh" các đoạn âm thanh, làm cho câu nói bị ngắt quãng, không liền mạch. Điều này xảy ra có thể bởi vì CNN không có thông tin hồi quy từ trước như trong mô hình CNN-BiLSTM, thứ cho mô hình thông tin của trạng thái trước đó.

V. Kết luận

Bài báo đã giới thiệu về mô hình kết hợp CNN-BiLSTM dùng trong việc phát hiện giọng nói. Tuy chỉ được cấu hình đơn giản và không quá tốn thời gian huấn luyện, mô hình vẫn cho kết quả đáng khả quan trên tập tiếng Việt, hứa hẹn khả năng mở rộng và cải tiến. Mô hình là bước tiền xử lý cho nhiều hệ thống: xử lý giọng nói, nhận dạng tiếng nói, v.v. được sử dụng trong những hệ thống trợ lý ảo, do đó hướng nghiên cứu tiếp theo ngoài cải thiện độ chính xác, mô hình có thể học được nhận dạng đặc trưng của người nói như giới tính, vùng miền, ngôn ngữ, và hơn thế nữa.

Cảm tạ

Bài báo này được tài trợ bởi Công ty Cổ phần Công nghệ OLLI, thực hiện trong quá trình sinh viên thực tập tại công ty.

Tài liệu

[1] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, "Applying support vector machines to voice activity detection," in *6th International Conference on Signal Processing, 2002.*, vol. 2, 2002, pp. 1124–1127 vol.2.

[2] *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*. ISCA, 2006.

[3] A. Amehraye-Fillatre, L. Fillatre, and N. Evans, "Voice activity detection based on a statistical semi-parametric test," 05 2013.

[4] Y. ZHANG, Z.-M. Tang, Y.-P. Li, and Y. LUO, "Enhanced voice activity detection using modified wiener filtering and harmonic structure information," *Sensors and Transducers*, vol. 166, pp. 275–280, 03 2014.

[5] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.

[6] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on youtube using deep neural networks," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 728–731, 01 2013.

[7] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7378–7382.

[8] X.-L. Zhang and D. Wang, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 252–264, 2016.

[9] A. Sehgal and N. Kehtarnavaz, "A convolutional neural network smartphone app for real-time voice activity detection," *IEEE Access*, vol. 6, pp. 9017–9026, 2018.

[10] S. Mihalache and D. Burileanu, "Using voice activity detection and deep neural networks with hybrid speech feature extraction for deceptive speech detection," *Sensors*, vol. 22, p. 1228, 02 2022.

[11] D. Augusto, J. Stuchi, R. Violato, and L. Cuozzo, *Exploring Convolutional Neural Networks for Voice Activity Detection*, 07 2017, pp. 37–47.

[12] T. Alam and A. Khan, "Lightweight cnn for robust voice activity detection," in *SPECOM*, 2020.

[13] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2519–2523, 2014.

[14] N. Wilkinson and T. Niesler, "A hybrid cnn-bilstm voice activity detector," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6803–6807.

[15] AILab. Vivos corpus. [Online]. Available: https://ailab.hcmus.edu.vn/vivos

[16] Y. Obuchi, "Framewise speech-nonspeech classification by neural networks for voice activity detection with statistical noise suppression," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5715–5719.

[17] M. H. Md Shahrin, "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks," 06 2017.