ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN
BỘ MÔN KHOA HỌC MÁY TÍNH

SINH TRẮC HỌC GIỌNG NÓI
VOICE BIOMETRICS

Nguyễn Hoàng Đức, Lê Nhựt Nam, Nguyễn Viết Dũng

GV Lý thuyết: **PGS. TS** Lê Hoàng Thái
GV Hướng dẫn: Nguyễn Ngọc Thảo, Lê Thanh Phong

Ngày 8 tháng 6 năm 2021

Nội dung trình bày

A. Trình bày nội dung tìm hiểu được từ Chapter 8 - Voice Biometrics

- Giới thiệu
- Xác định những thông tin trong tín hiệu giọng nói
- Rút trích đặc trưng và Phân tách dữ liệu
- Nhận dạng giọng nói phụ thuộc văn bản
- Nhận dạng giọng nói không phụ thuộc văn bản
- Ứng dụng

Nội dung trình bày

B. Trình bày các phương pháp STATE OF THE ART của Voice Recognition

- Động lực nghiên cứu khoa học
- Kho ngữ liệu/ Cơ sở dữ liệu
- Phát biểu bài toán
- Các công trình liên quan
- Demo
- Tài liệu tham khảo

Giới thiệu

- Giọng nói (Voice/Speech) là một đặc điểm sinh trắc học (nhân trắc học) dễ dàng tiếp cận nhất mà không cần phải có thêm thiết bị thu nhận và hệ thống truyền dẫn.
- Có lợi thế khi áp dụng vào các hệ thống điều khiển từ xa
- Giọng nói không chỉ liên quan đến các đặc trưng cá thể mà còn liên quan với môi trường xung quanh và vấn đề xã hội, do vậy việc sản sinh giọng nói là một kết quả của một quá trình hết sức phức tạp.

Những thông tin nhận dạng trong tín hiệu giọng nói

- Idiolectal characteristics: cách phát âm phản ánh khu vực bạn đang sống hoặc đã sống và các phong cách nói khác nhau thay đổi một cách tinh vi tùy thuộc vào người bạn đang nói đến.
- Phonotactics characteristics: Mô tả cách sử dụng của người nói của các đơn vị ngữ âm và khả năng nhận ra khả dụng.
- Prosody characteristics: Prosody (ngữ điệu), là sự kết hợp của năng lượng tức thời, âm điệu, tốc độ nói và thời lượng đơn vị để cung cấp cho lời nói sự tự nhiên, đầy đủ ý nghĩa và giọng điệu cảm xúc.
- Short-term spectral characteristics: liên quan trực tiếp với hành động phát âm đơn lẻ, quan hệ với tạo ra ngữ âm và cũng liên quan đến sinh lý cá nhân của quá trình phát sinh giọng nói. Đây là một thông tin nhận dạng quan trọng

Phân tích cửa sổ (Short-term Analysis)

Các tín hiệu giọng nói thường thay đổi liên tục nên người ta thường giới hạn cửa sổ phân tích, thường dùng cửa sổ dạng cosine như hamming hoặc hanning, với độ dài ngắn từ 20 đến 40 mili giây, thường được gọi là tín hiệu giả tĩnh trên mỗi khung.

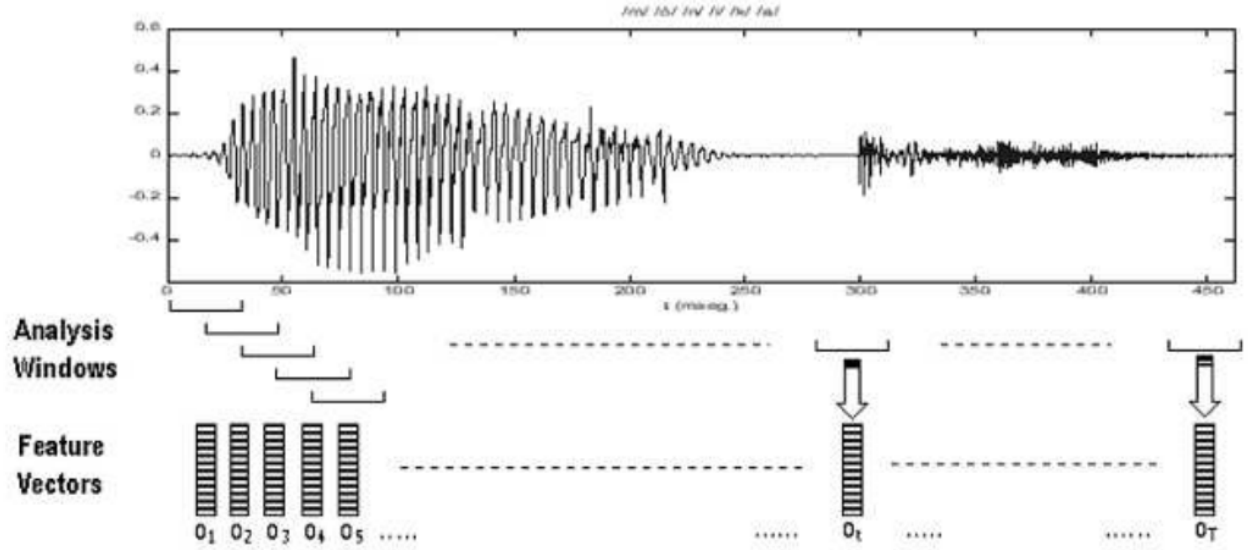
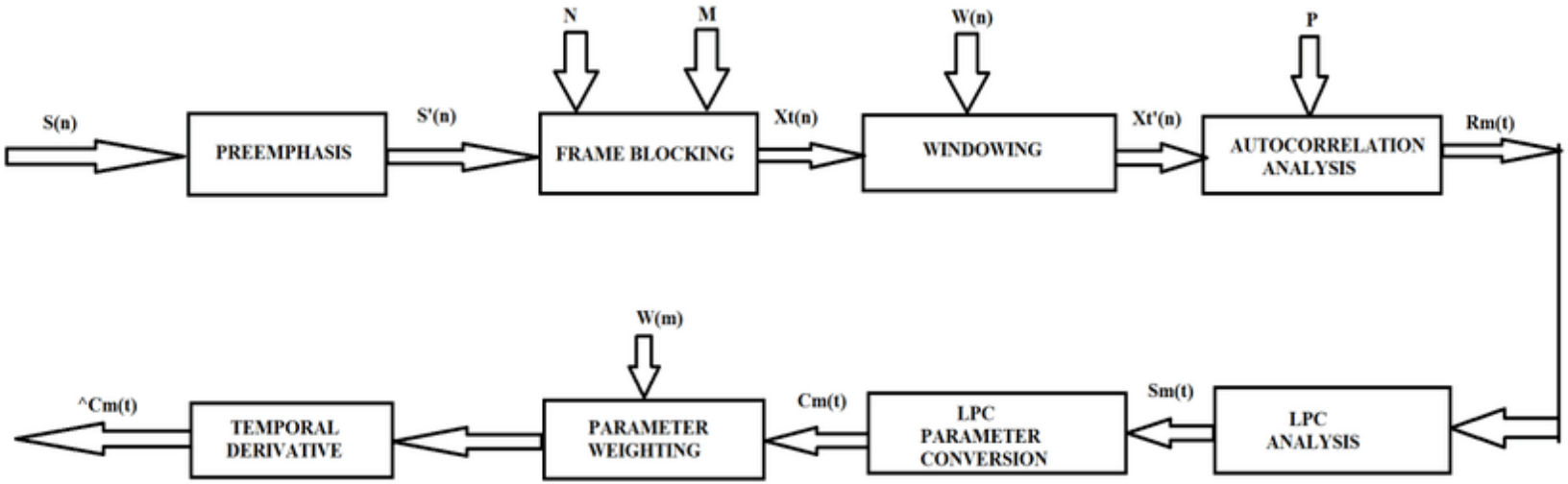


Fig. 8.1. Short-term analysis and parameterization of a speech signal.

Hình 1: Handbook of Biometrics, page 155

Tham số hóa

Tham số hóa bằng cách dùng Linear Predictive Coding (LPC): là một phương pháp được sử dụng hầu hết trong xử lý tín hiệu âm thanh và xử lý giọng nói để biểu diễn đường bao phổ của tín hiệu số của tiếng nói ở dạng nén, sử dụng thông tin của mô hình dự đoán tuyến tính.

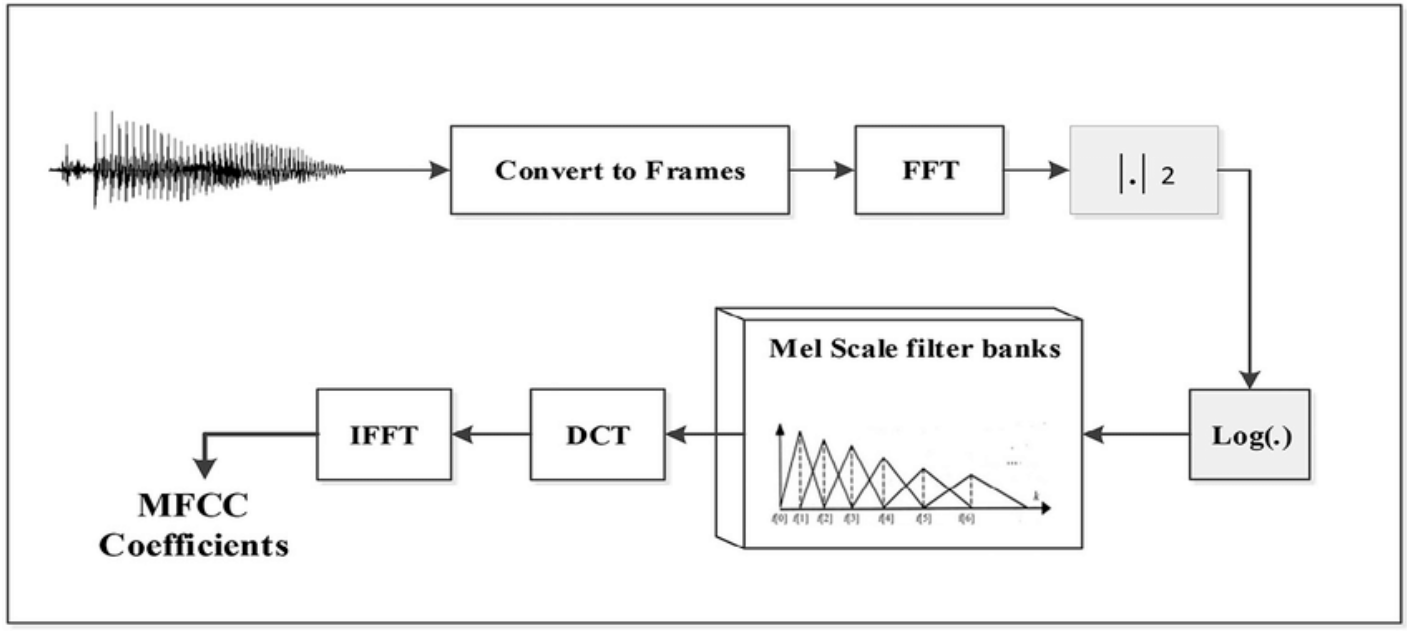


Hình 2: Handbook of Biometrics, page 162

Tham số hóa

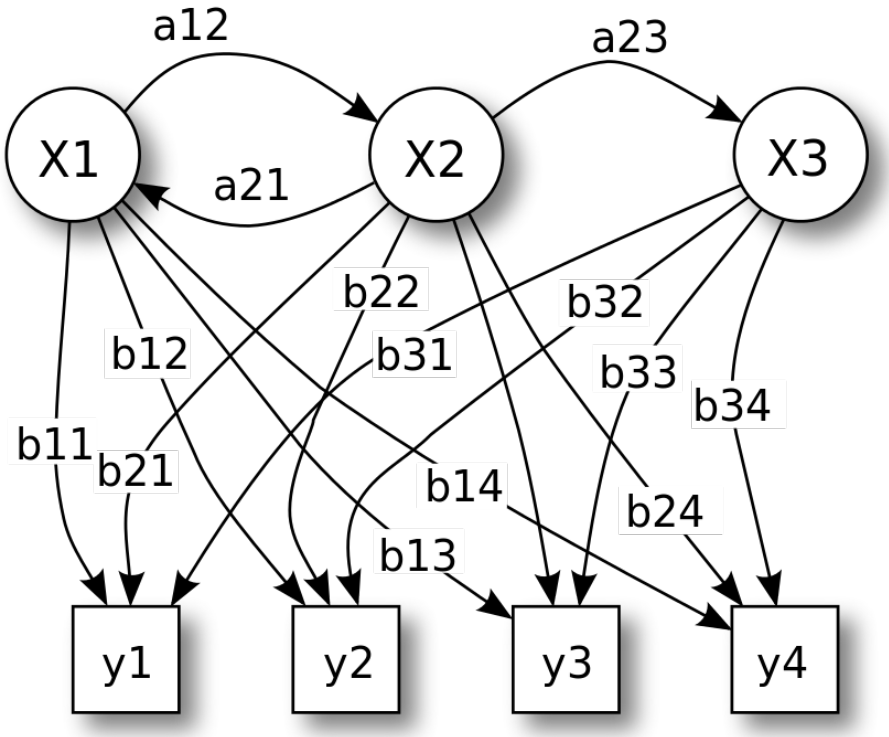
Tham số hóa bằng cách dùng Mel-Frequency based Cepstral Coefficients (MFCC): MFCC là một kỹ thuật rút trích đặc trưng giọng nói được dùng nhiều trong các mô hình nhận dạng giọng nói, nó cho ra các kết quả là các hệ số (Coefficients) của cepstral từ Mel Filter trên phổ từ dữ liệu âm thanh giọng nói.

$$\text{speech signal} \Rightarrow \text{spectrum} \Rightarrow \text{mel-freq filter} \Rightarrow \text{cepstral}$$



Phân tích ngữ âm và tách từ

Mô hình Hidden Markov (HMM - Hidden Markov Models) là công cụ thành công nhất và được sử dụng rộng rãi (ngoại trừ một số kiến trúc ANN) để mã hóa ngữ âm, âm tiết và từ, nghĩa là dịch từ lời nói được lấy mẫu thành một căn chỉnh thời gian dãy các đơn vị ngôn ngữ.



Phân tách ngữ điệu

Dựa trên cơ sở là cao độ và năng lượng ở từng frame

- Cao độ: xác định bằng phương pháp tự động tương quan, phân rã cepstral dựa trên một số phương thức làm mịn bằng bộ lọc.
- Năng lượng: Năng lượng cửa sổ thu được rất dễ dàng thông qua định lý Parseval.

Giới thiệu

Công nghệ nhận dạng giọng nói phụ thuộc văn bản, sử dụng nội dung từ vựng của giọng nói phát ra để nhận dạng giọng nói, ứng dụng chính của hệ thống này trong các hệ thống tương tác, nơi cần có sự hợp tác từ người dùng để xác thực danh tính của họ.
Phân loại

- Hệ thống văn bản tĩnh: nội dung từ vựng trong ghi danh và các mẫu nhận dạng luôn giống nhau.
- Hệ thống văn bản động: tạo ra một lời nhắc mật khẩu được tạo ngẫu nhiên khác nhau mỗi khi người dùng được xác minh (hệ thống nhắc bằng văn bản)

Phương pháp thực hiện

Có hai phương pháp thường được dùng:

- Phương pháp dựa trên khuôn mẫu: bao gồm một số chuỗi vector tương ứng với lời nói đăng ký và việc nhận dạng được thực hiện bằng cách so sánh lời nói xác minh với lời nói đăng ký.
- Phương pháp thống kê: Nổi bật nhất là mô hình Markov ẩn (HMM), cho phép chọn đơn vị tiếng nói từ đơn vị âm vị phụ đến từ và cho phép thiết kế hệ thống nhắc văn bản.

Giới thiệu

Công nghệ nhận dạng giọng nói độc lập với văn bản cố gắng giảm thiểu ảnh hưởng của nội dung từ vựng vốn được coi là không xác định đối với khả năng nhận dạng của giọng nói, điều này trái ngược với hệ thống nhận dạng giọng nói phụ thuộc văn bản, đương nhiên việc nghiên cứu và phát triển nó sẽ khó khăn hơn.

- Hệ thống cửa sổ phổ âm
- Hệ thống Idiolectal
- Hệ thống ngữ âm
- Hệ thống ngữ điệu

Hệ thống cửa sổ phổ âm

Dùng trong việc phân tích phổ trong khoảng thời gian ngắn được sử dụng để mô hình các đặc trưng người nói, nhằm mô hình hóa các "âm thanh" khác nhau mà một người có thể tạo ra

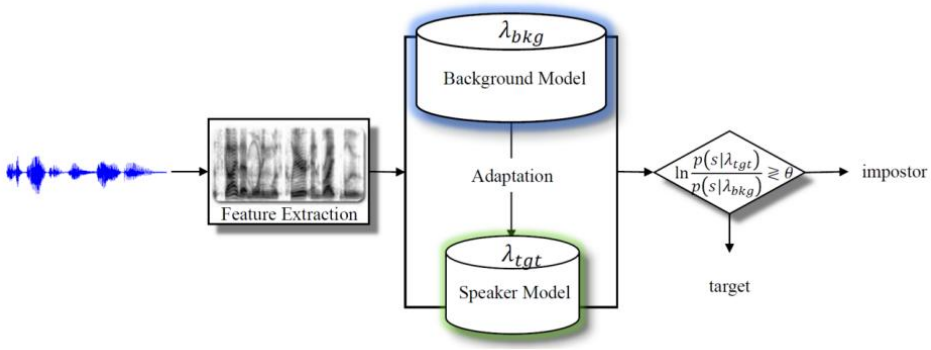
- Kỹ thuật Lượng tử hóa Vector (Vector Quantization techniques)
- Gaussian Mixture Model - Universal Background Model
- Discriminative techniques - Kỹ thuật phân tách: Artificial Neural Networks, SVM - Support Vector Machine
- SuperVectors, một kỹ thuật hỗn hợp GMM-SVM

Hệ thống Idiolectal

- Doddington đã lập ra mô hình cách sử dụng từ của từng người nói cụ thể bằng cách sử dụng **n-gram** mô hình hóa các chuỗi từ và xác suất của chúng và chứng minh rằng việc sử dụng các mô hình đó có thể cải thiện hiệu suất của hệ thống GMM âm thanh/ phổ cơ bản.
- Quan trọng hơn kết quả cụ thể này là thực tế là công trình này đã thúc đẩy nghiên cứu trong việc sử dụng các cấp độ thông tin cao hơn (idiolectal, phonotactic, prosodic, v.v.) để **nhận dạng giọng nói độc lập với văn bản**.

Hệ thống ngữ âm

Mô hình hệ thống ngữ âm



Hình 4: Quy trình của hệ thống ngữ âm

- Bước 1: Đầu vào là giọng nói cần xác minh
- Bước 2 - Token segmentation: Tách thành các đoạn t3, t4, t5, t3
- Bước 3: Mô hình hóa ngôn ngữ thống kê n-gram
  - Huấn luyện mô hình Universal Background Phone (UBPM) $L_U = P(X|UBPM)$
  - Huấn luyện mô hình Speaker Phone Models (SPM_i): $L_{S_i} = P(X|SPM_i)$
- Bước 4: Tính recognition score: $Score_i = \frac{1}{m} log\left(\frac{P(X|SPM_i)}{P(X|UBPM)}\right)$

Hệ thống ngữ điệu

Mô hình hệ thống ngữ điệu

- Giai đoạn một, đối với mỗi đoạn giọng nói của đoạn thoại, quỹ đạo thời gian của các đặc điểm ngữ điệu (tần số cơ bản - hoặc cao độ- và năng lượng) được rút trích.
- Giai đoạn hai, cả hai đường bao đều được phân đoạn và dán nhãn bằng định lượng trung bình độ dốc.

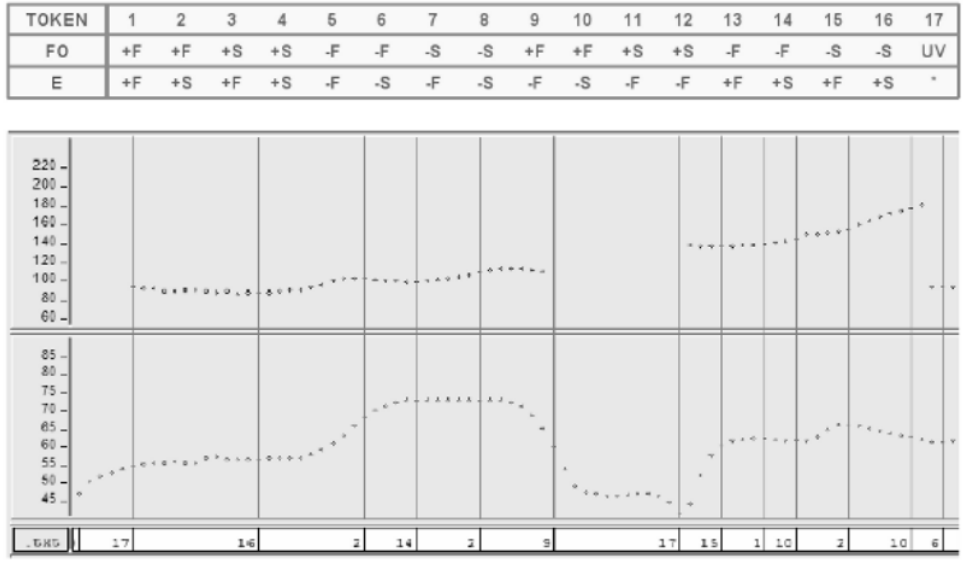
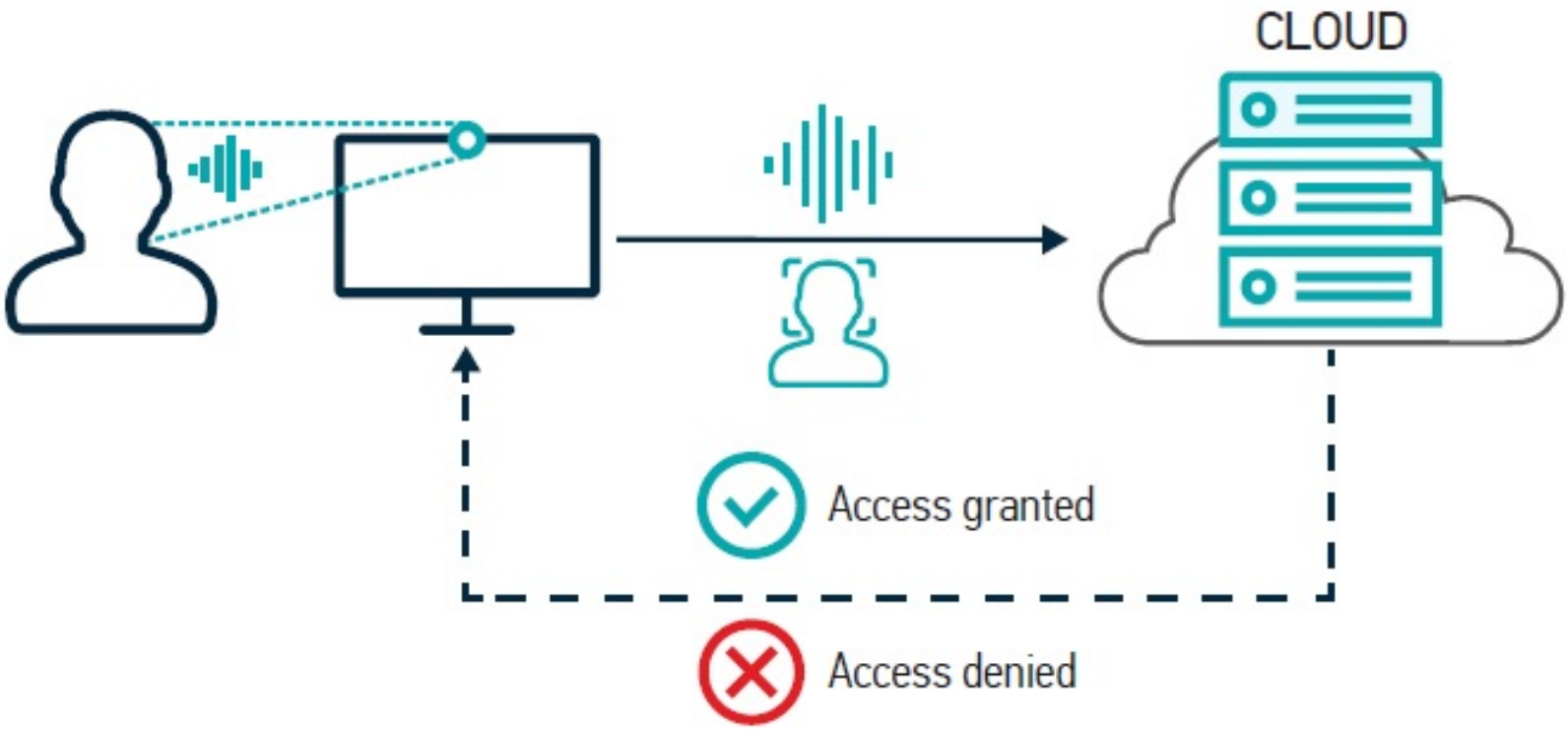


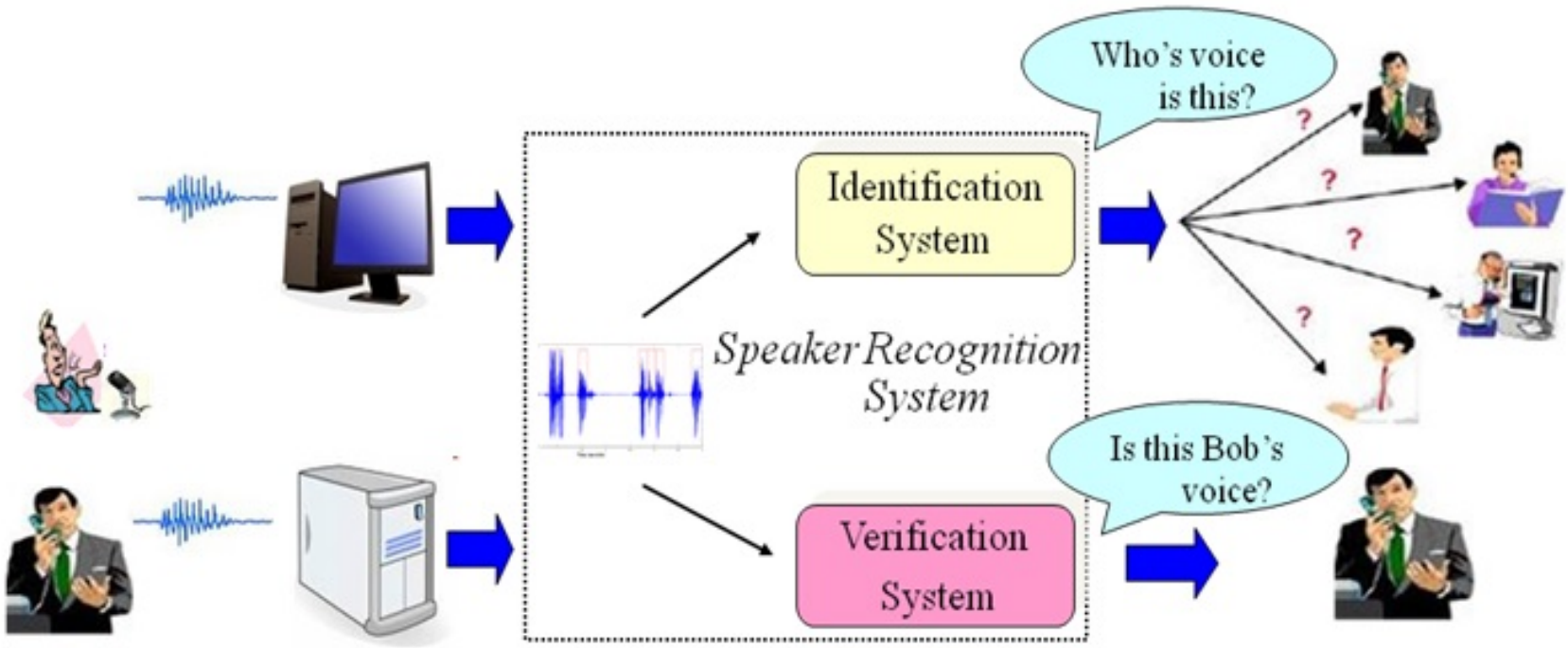
Fig. 8.4. Prosodic token alphabet (top table) and sample tokenization of pitch and energy contours (bottom figure).

Voice authentication



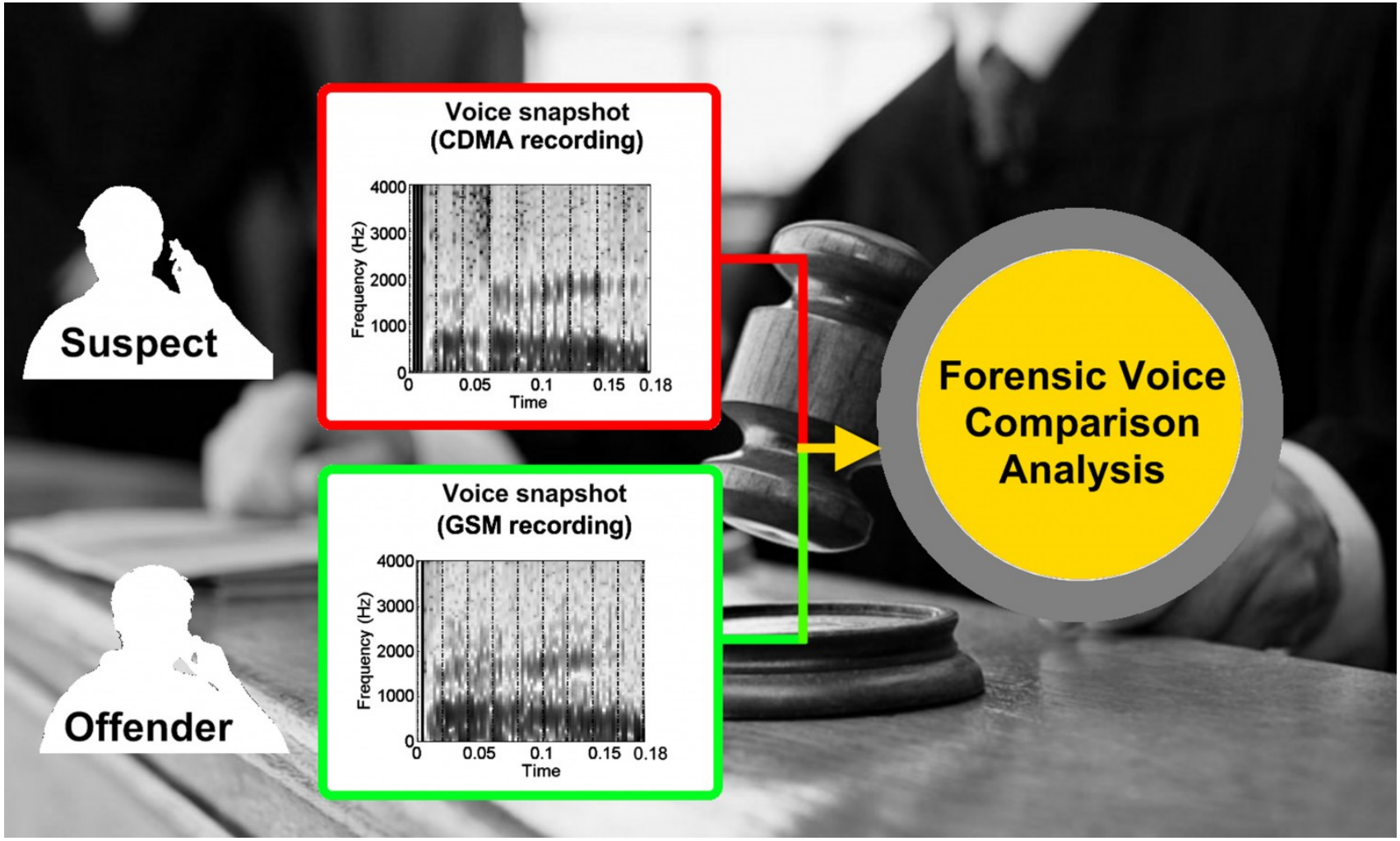
Hình 6: Ví dụ Voice authentication/ Verification

Speaker Identification and Verification



Hình 7: Ví dụ Speaker Recognition Systems

Forensic speaker recognition



Hình 8: Ví dụ Speaker Recognition Systems

Động lực nghiên cứu khoa học

- Những phương pháp đã tìm hiểu từ sách Handbook of Biometrics: Voice Biometrics đã cho chúng ta cái nhìn tổng quan về lĩnh vực Nhận dạng giọng nói và những phương pháp truyền thống (tạm gọi là thời kỳ trước Deep Learning) cùng với những thông tin các công trình nghiên cứu nổi bật.
- Các phương pháp SOTA dựa trên việc biểu diễn i-vectors của những đoạn giọng nói, cải thiện đáng kể so với mô hình Gaussian Mixture Model-Universal Background Models
- Sự phát triển của Deep Learning

Một số kho ngữ liệu lớn, nổi bật gần đây

Kho ngữ liệu VCTK (Veaux et al., 2017)

- Tên đầy đủ: SUPERSEDED - CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit
- Tác giả/ Nhóm tác giả: Veaux Christophe, Yamagishi Junichi, MacDonald Kirsten
- Nhà xuất bản: University of Edinburgh. The Centre for Speech Technology Research (CSTR)
- Mô tả sơ lược: Tất cả dữ liệu giọng nói được ghi lại bằng cách sử dụng thiết lập ghi âm giống hệt nhau: micrô đa hướng (DPA 4035) và micrô tụ màng nhỏ với băng thông rất rộng (Sennheiser MKH 800), tần số lấy mẫu 96kHz ở 24 bit và trong một buồng phản xạ hemi của Đại học Edinburgh

Một số kho ngữ liệu lớn, nổi bật gần đây

Kho ngữ liệu VoxCeleb (Nagrani et. al, 2017), VoxCeleb2 (Chung et. al, 2018)

- Tên đầy đủ: VoxCeleb2: Deep Speaker Recognition
- Tác giả/ Nhóm tác giả: J. S. Chung*, A. Nagrani*, A. Zisserman
- Hội nghị: INTERSPEECH, 2018.
- Mô tả sơ lược: Là kho ngữ liệu giọng nói lớn nhất hiện tại, VoxCeleb2 chứa hơn 1 triệu câu nói cho 6.112 người nổi tiếng, được trích xuất từ các video tải lên YouTube. Tập hợp phát triển của VoxCeleb2 không có sự trùng lặp với các đặc điểm nhận dạng trong tập dữ liệu VoxCeleb1 hoặc SITW.

Phát biểu bài toán

Tác vụ: Định danh người nói (Speaker Identification)

- Đầu vào (Input): Dữ liệu âm thanh giọng nói (Sound)
- Đầu ra (Output): Danh tính của người nói (Personal Information, Label)

Tác vụ: Xác nhận người nói (Speaker Verification)

- Đầu vào (Input): Dữ liệu âm thanh giọng nói (Sound)
- Đầu ra (Output): Đồng ý/ Từ chối (Accept/ Deny)

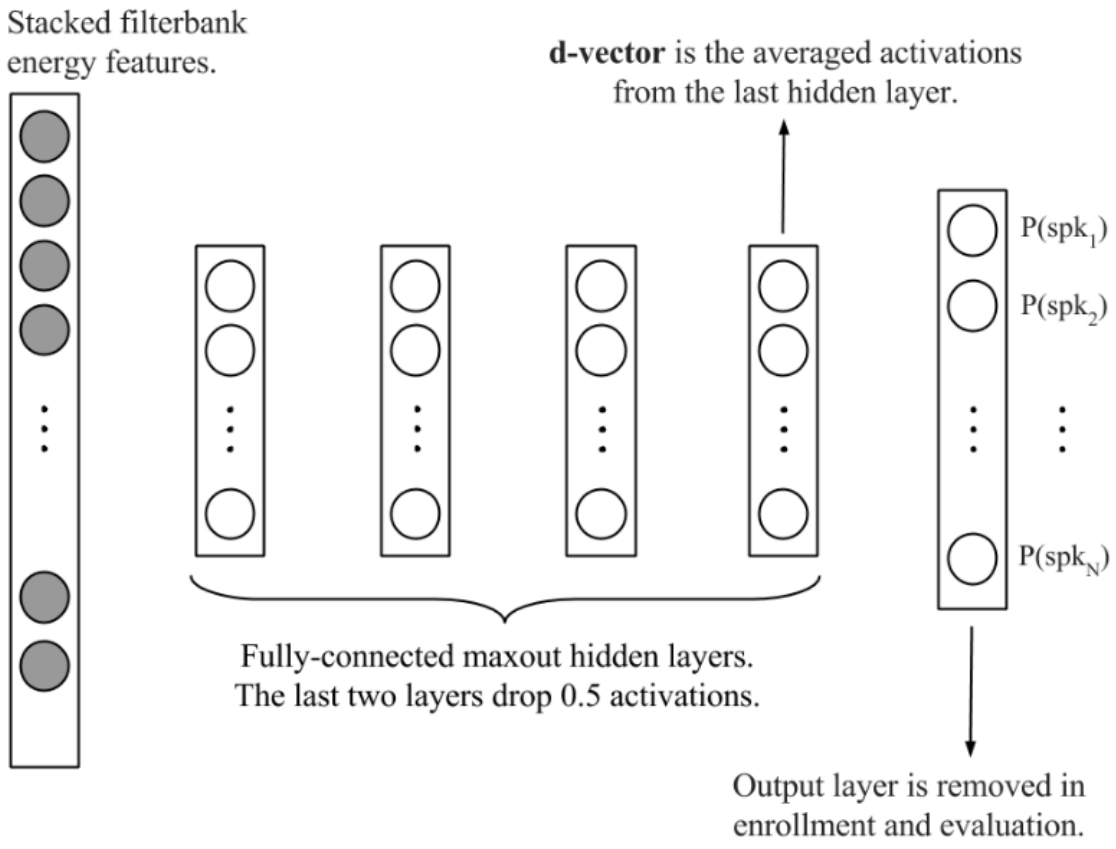
Công trình tiêu biểu sử dụng d-vectors

Được giới thiệu trong bài báo "Deep neural networks for small footprint text-dependent speaker verification"**d-vectors** trở thành tiền đề cho hàng loạt các thành công sau này của lĩnh vực Nhận dạng giọng nói sử dụng Deep Learning. **Giới thiệu chung về bài báo:**

- Bài báo: Deep neural networks for small footprint text-dependent speaker verification (Một bước nhỏ trong dùng mạng học sâu cho tác vụ xác minh người nói)
- Nhóm tác giả: Ehsan Variani (Johns Hopkins Univ., Baltimore, MD, USA), Xin Lei, Erik McDermott, Ignacio Lopez Moreno, Javier Gonzalez-Dominguez (Google Inc., USA)
- Được publish tại hội nghị 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), diễn ra tại Florence, Italy, vào năm 2014

Công trình tiêu biểu sử dụng d-vectors

Mô hình d-vectors



Hình 9: Mô hình DNN với d-vectors

Công trình tiêu biểu sử dụng d-vectors

Phương pháp tiếp cận

- Sử dụng Deep Neural Networks trong tác vụ xác minh người nói (Speaker Verification) để rút trích đặc trưng giọng nói và kích hoạt đặc trưng nằm lớp ẩn cuối cùng của mạng học - gọi là d-vectors (Deep Vectors)
- Trong giai đoạn đăng ký (Speaker Enrollment), mô hình DNN được huấn luyện sẵn được sử dụng để rút trích đặc trưng giọng nói của người nói ở lớp cuối cùng. Sau đó tính toán giá trị trung bình của những đặc trưng này, để cho ra d-vector của người nói
- Trong giai đoạn đánh giá (Evaluation Stage), một giọng nói cần được xác minh khi vào mô hình sẽ được tính toán để cho ra một d-vector. Sau đó, dùng d-vector này so sánh với những d-vector trong quá trình đăng ký để đưa ra quyết định giọng nói của người nói này có phải là đúng với danh tính đưa ra không.

Công trình tiêu biểu sử dụng d-vectors

Kết quả đạt được

Table 2. EER results of i-vector and d-vector verification systems using different number of utterances for enrollment.

	# utterances in enrollment			
	4	8	12	20
i-vector	2.83%	2.06%	1.64%	1.21%
d-vector	4.54%	3.21%	2.64%	2.00%

Hình 10: - Deep Neural Networks for small foot-print text-dependent speaker verification - 2014

Nhận xét:

- Cơ sở phương pháp, độ đo đánh giá đơn giản
- EER vẫn chưa tốt so với hệ thống i-vectors

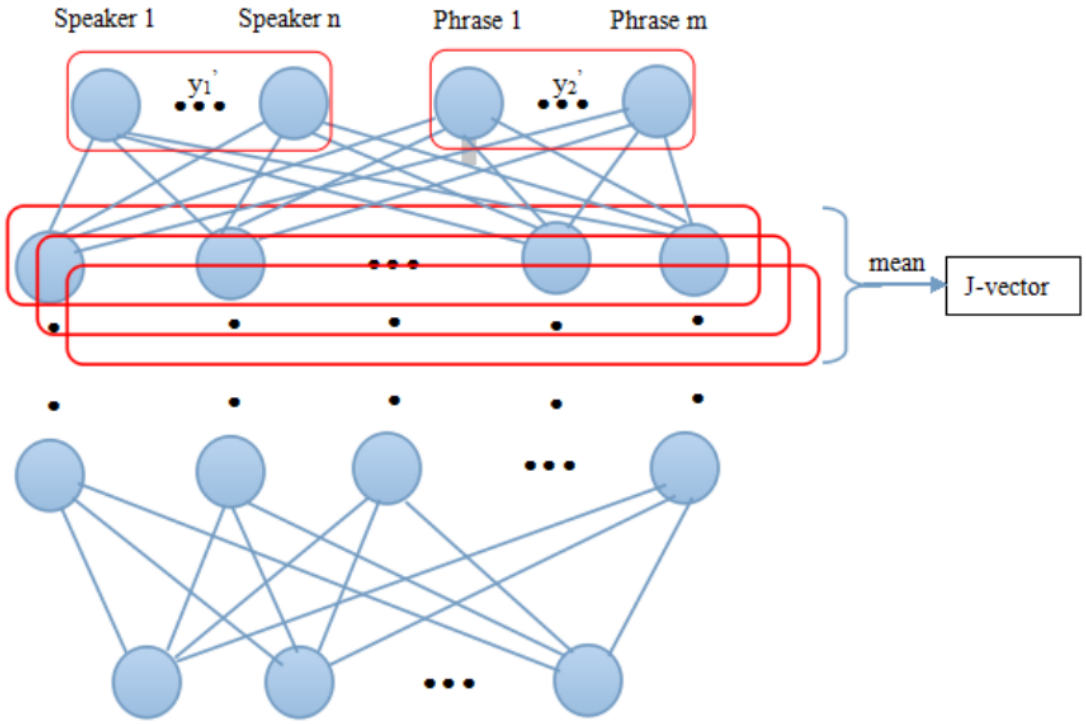
Công trình tiêu biểu sử dụng j-vectors

Giới thiệu chung về bài báo:

- Bài báo: Multi-Task Learning for Text-Dependent Speaker Verification
- Nhóm tác giả: Nanxin Chen, Yanmin Qian, Kai Yu (Shanghai Jiao Tong University, China)
- Được publish tại hội nghị INTERSPEECH 2015 16th Annual Conference of the International Speech Communication Association, diễn ra tại Dresden, Germany, từ ngày 6-10 tháng 9 năm 2015

Công trình tiêu biểu sử dụng j-vectors

Mô hình j-vectors



Hình 11: Mô hình DNN với j-vectors

Công trình tiêu biểu sử dụng j-vectors

Phương pháp thực hiện: Mô hình Học đa nhiệm Deep Neural Network sử dụng cho Xác minh người nói phụ thuộc văn bản được lấy ý tưởng từ việc sử dụng DNN với số lượng tham số cực lớn, một mô hình DNN có thể học cùng lúc việc phân tách văn bản, lẫn giọng nói
Hai hàm mất mát ban đầu là $C_1(y_1, y_1^{'}), C_2(y_2, y_2^{'})$ được dùng để tạo thành hàm tổng mất mát:

$$C([y_1, y_2], [y_1^{'}, y_2^{'}]) = C_1(y_1, y_1^{'}) + C_2(y_2, y_2^{'})$$

Trong đó:

- C_1, C_2 lần lượt là hai cross-entropy criteria cho giọng nói và văn bản
- y_1, y_2 đại diện cho nhãn đúng của từng người nói và văn bản
- $y_1^{'}, y_2^{'}$ đại diện cho nhãn đầu ra (nhãn dự đoán được) của y_1, y_2 tương ứng

Công trình tiêu biểu sử dụng j-vectors

Các kết quả đạt được

Table 2: Performance for different deep learning systems

Feature	Classifier	EER	minDCF
r-vector	Cosine Sim.	17.43	0.684
	Joint GDF	0.80	0.037
	Joint PLDA	1.47	0.065
d-vector	Cosine Sim.	21.05	0.818
	Joint GDF	0.71	0.033
	Joint PLDA	1.62	0.070
j-vector	Cosine Sim.	9.85	0.466
	Joint GDF	**0.14**	**0.007**
	Joint PLDA	0.54	0.027

Table 3: Performance under unseen speakers conditions

Unseen Speakers Ratio		1/5		1/3	
Feature	Classifier	EER	minDCF	EER	minDCF
r-vector	Cosine Sim.	20.68	0.820	20.71	0.818
	Joint GDF	1.33	0.062	1.63	0.076
	Joint PLDA	1.42	0.066	1.65	0.073
d-vector	Cosine Sim.	15.75	0.644	15.75	0.654
	Joint GDF	1.43	0.063	1.78	0.079
	Joint PLDA	1.56	0.063	1.65	0.067
j-vector	Cosine Sim.	9.65	0.463	9.64	0.464
	Joint GDF	**0.47**	0.033	0.58	0.050
	Joint PLDA	0.50	**0.022**	**0.50**	**0.024**

Nhận xét:

- Áp dụng học đa nhiệm (giọng nói và văn bản), rút trích ra j-vectors
- Kết quả tốt hơn, cho độ lỗi thấp hơn rất nhiều so với d-vector, r-vector trên tác vụ xác minh người nói phụ thuộc văn bản

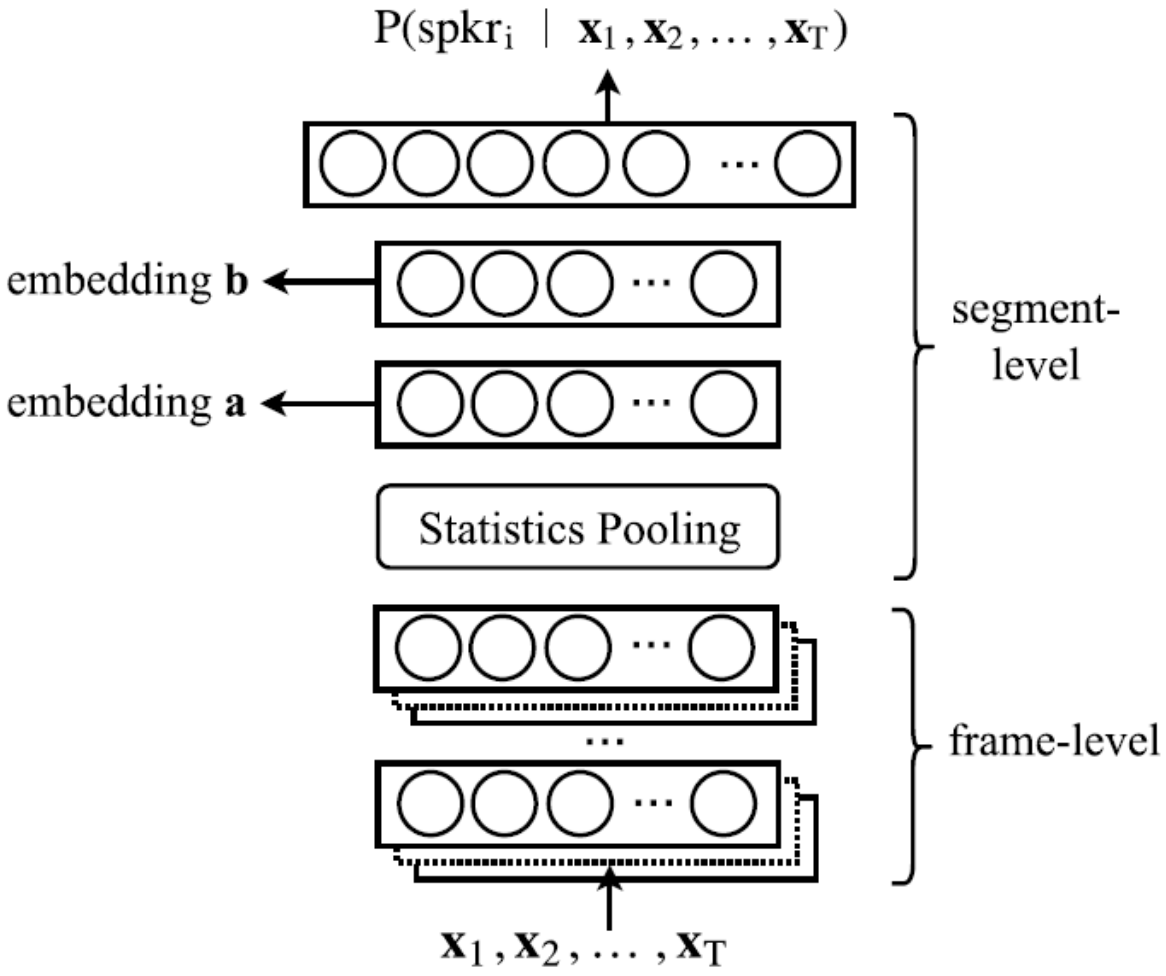
Công trình tiêu biểu sử dụng x-vectors

Giới thiệu chung về bài báo:

- Bài báo: X-Vectors: Robust DNN Embeddings for Speaker Recognition
- Nhóm tác giả:
  - David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, Sanjeev Khudanpur - Center for Language and Speech Processing & Human Language Technology Center of Excellence, The Johns Hopkins University, Baltimore, MD, USA
- Được xuất bản tại hội nghị 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), diễn ra tại Calgary, AB, Canada, năm 2018

Công trình tiêu biểu sử dụng x-vectors

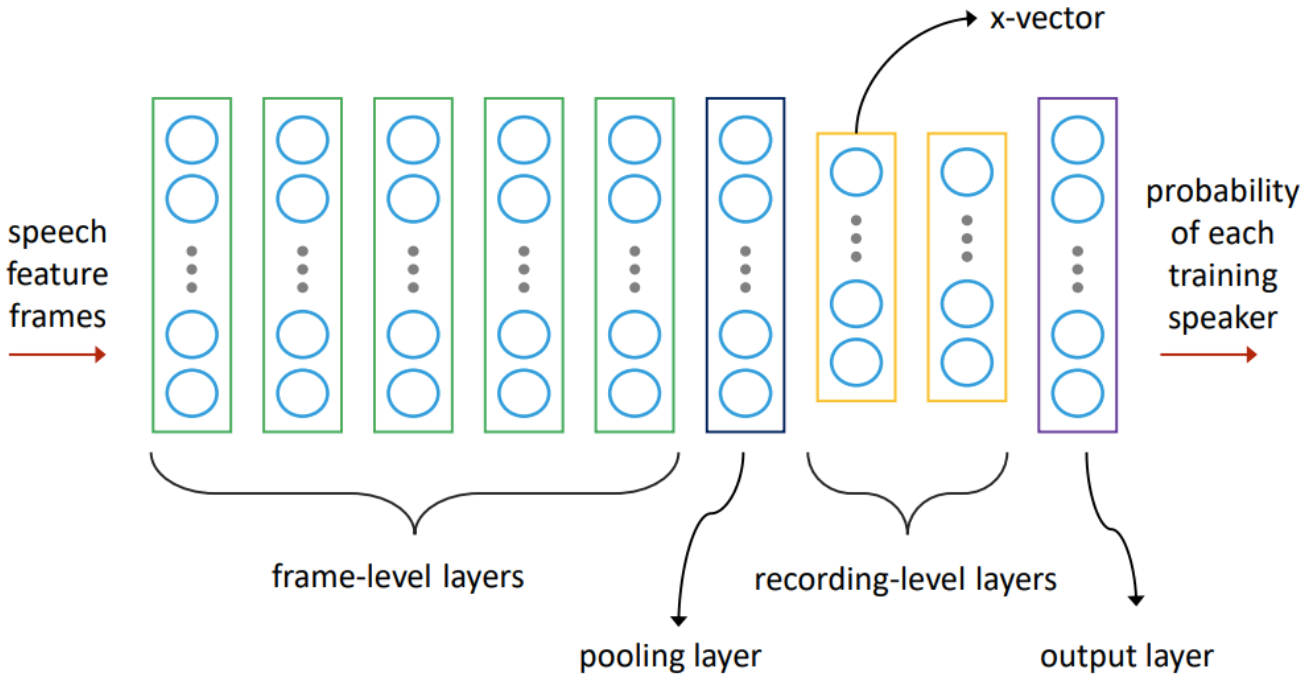
Mô hình x-vectors



Hình 12: Mô hình DNN với x-vectors

Công trình tiêu biểu sử dụng x-vectors

Phương pháp thực hiện



Hình 13: From i-vectors to x-vectors – a generational change in speaker recognition illustrated on the NFI-FRIDA database - OxfordWave Research

x-vectors sẽ được rút trích tại lớp phân đoạn 6 (segment6), ngay sau khi qua lớp tổng hợp thống kê

Công trình tiêu biểu sử dụng x-vectors

Các kết quả đạt được

			SITW Core			SRE16 Cantonese		
			EER(%)	$DCF10^{-2}$	$DCF10^{-3}$	EER(%)	$DCF10^{-2}$	$DCF10^{-3}$
4.1	Original systems	i-vector (acoustic)	9.29	0.621	0.785	9.23	0.568	0.741
		i-vector (BNF)	**9.10**	**0.558**	**0.719**	9.68	0.574	0.765
		x-vector	9.40	0.632	0.790	**8.00**	**0.491**	**0.697**
4.2	PLDA aug.	i-vector (acoustic)	8.64	0.588	0.755	8.92	0.544	0.717
		i-vector (BNF)	8.00	**0.514**	**0.689**	8.82	0.532	0.726
		x-vector	**7.56**	0.586	0.746	**7.45**	**0.463**	**0.669**
4.3	Extractor aug.	i-vector (acoustic)	8.89	0.626	0.790	9.20	0.575	0.748
		i-vector (BNF)	7.27	**0.533**	0.730	8.89	0.569	0.777
		x-vector	**7.19**	0.535	**0.719**	**6.29**	**0.428**	**0.626**
4.4	PLDA and extractor aug.	i-vector (acoustic)	8.04	0.578	0.752	8.95	0.555	0.720
		i-vector (BNF)	6.49	0.492	0.690	8.29	0.534	0.749
		x-vector	**6.00**	**0.488**	**0.677**	**5.86**	**0.410**	**0.593**
4.5	Incl. VoxCeleb	i-vector (acoustic)	7.45	0.552	0.723	9.23	0.557	0.742
		i-vector (BNF)	6.09	0.472	0.660	8.12	0.523	0.751
		x-vector	**4.16**	**0.393**	**0.606**	**5.71**	**0.399**	**0.569**

Table 2. Results using data augmentation in various systems. "Extractor" refers to either the UBM/**T** or the embedding DNN. For each experiment, the best results are **boldface**.

Hình 14: x-vector DNN embedding architecture in (Snyder et al., 2018)

Nhận xét

- Hiệu suất của x-vectors đã được chứng minh là tốt hơn đáng kể so với i-vectors, đặc biệt là ở khoảng thời gian ngắn.

So sánh d-vectors, j-vectors và x-vectors

	d-vectors	j-vectors	x-vector
Kỹ thuật rút trích	DNN	DNN	DNN
Ví trí rút trích	Tại lớp ẩn cuối cùng DNN	Tại lớp ẩn cuối cùng DNN	Sau khi qua lớp statistics pooling
Cách rút trích	Là trung bình kích hoạt tại lớp ẩn cuối cùng	Là trung bình kích hoạt tại lớp ẩn cuối cùng, kết hợp tín hiệu giọng nói và dữ liệu văn bản	Là vector phân đoạn (segment6) sau khi tính toán thống kê

Công trình tiêu biểu sử dụng Multi-domain features

Giới thiệu chung về bài báo:

- Tên bài báo: Multi-task Recurrent Model for Speech and Speaker Recognition
- Nhóm tác giả: Zhiyuan Tang (1) (2) , Lantian Li (1) and Dong Wang (1)
  - (1): Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology Center for Speech and Language Technologies, Research Institute of Information Technology, Tsinghua University
  - (2): Chengdu Institute of Computer Applications, Chinese Academy of Sciences
- Bài báo được đăng trên arXiv.org, vào ngày 31, tháng 3 năm 2016 (Phiên bản mới nhất được cập nhật vào ngày 27 tháng 9 năm 2016)

Công trình tiêu biểu sử dụng Multi-domain features

Mô hình

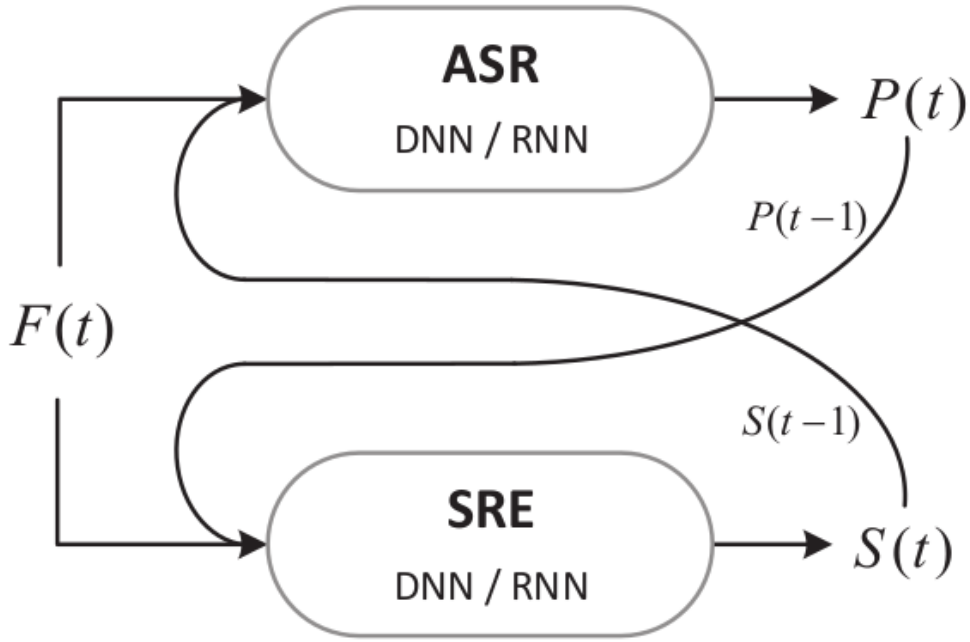


Fig. 1. Multi-task recurrent learning for ASR and SRE. $F(t)$ denotes primary features (e.g., Fbanks), $P(t)$ denotes phone identities (e.g., phone posteriors, high-level representations for phones), $S(t)$ denotes speaker identities (e.g., speaker posteriors, high-level representations for speakers).

Công trình tiêu biểu sử dụng Multi-domain features

Phương pháp thực hiện: Ý tưởng cơ bản của nó là sử dụng đầu ra của một tác vụ như một đầu vào vào của những tác vụ khác (khá tương tự với kiến trúc RNN thông thường). Kết quả đầu ra của một tác vụ ở bước thời gian trước đó $t-1$ được sử dụng để cung cấp cho một tác vụ ở thời điểm t hiện tại.

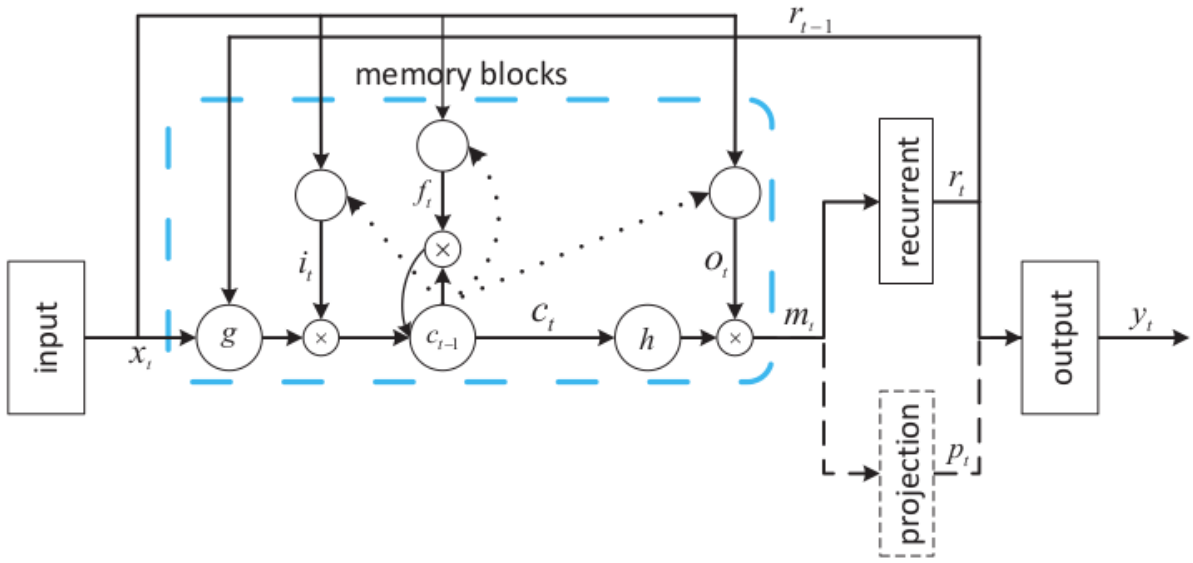


Fig. 2. Basic recurrent LSTM model for ASR and SRE single-task baselines. The picture is reproduced from [11].

Công trình tiêu biểu sử dụng Multi-domain features

Các kết quả đạt được

TABLE I
ASR BASELINE RESULTS.

	dev92	eval92	eval93	Total
WER%	8.36	5.14	8.06	7.41

TABLE II
SRE BASELINE RESULTS.

	EER%		
System	Cosine	LDA	PLDA
i-vector (200)	2.89	1.03	0.57
r-vector (256)	1.84	1.34	3.18

TABLE III
JOINT TRAINING RESULTS.

Feedback Info.		Feedback Input				ASR WER%	SRE EER%
r	p	i	f	o	g		
						7.41	1.84
√		√				7.05	0.62
√	√	√				6.97	0.64
√			√			7.12	0.66
√	√		√			7.24	0.65
√				√		7.26	0.65
√	√			√		7.28	0.59
√					√	7.11	0.62
√	√				√	7.11	0.67
√		√	√	√		7.06	0.66
√	√	√	√	√		7.23	0.71
√		√	√	√	√	7.05	0.55
√	√	√	√	√	√	7.23	0.62

Nhận xét

- Một kiến trúc mạng học lặp lại có thể huấn luyện cùng lúc nhiều tác vụ.
- Kết quả trên cơ sở dữ liệu cho thấy nhiều triển vọng cho kiến trúc nói riêng, và những hướng hướng tiếp cận đa nhiệm nói chung.

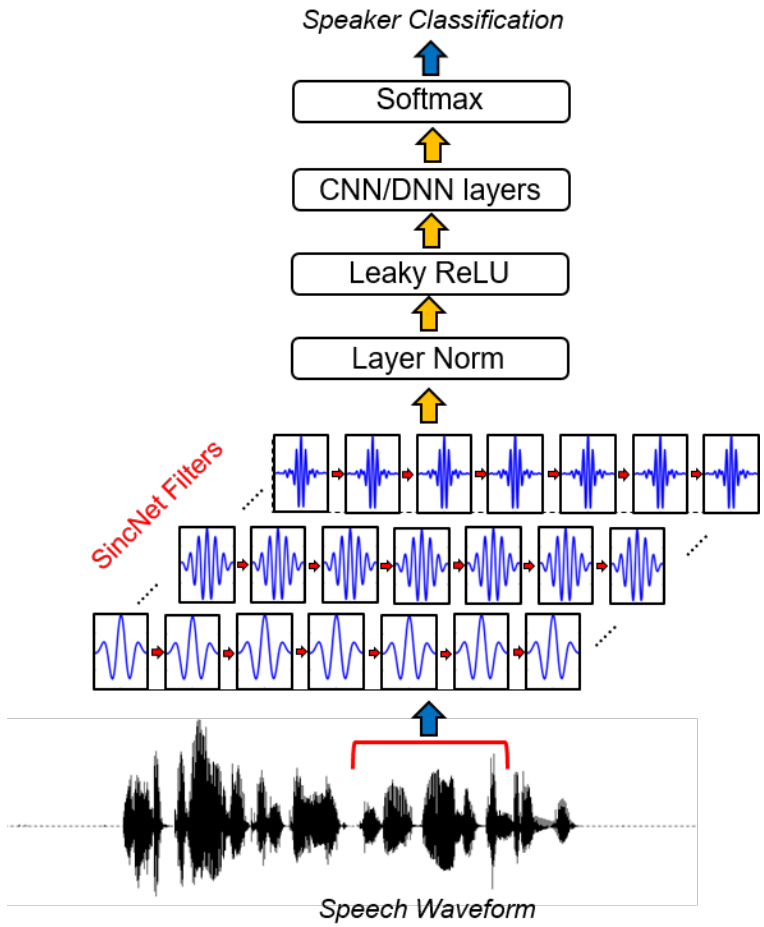
Công trình tiêu biểu: SincNet

Giới thiệu chung về bài báo:

- Bài báo: Speaker Recognition from Raw waveform with SincNet
- Nhóm tác giả: Mirco Ravanelli, Yoshua Bengio*
  - Mila, Université de Montréal
  - CIFAR Fellow*
- Được công bố trên arXiv.org vào năm 2018

Công trình tiêu biểu sử dụng SincNet

Mô hình



Hình 15: The SincNet Architecture

Công trình tiêu biểu: SincNet

Phương pháp thực hiện Sincnet thực hiện các phép tích chập của nó với hàm g, hàm này phụ thuộc vào một tham số θ. Công thức như sau:

$$y[n] = x[n] * g[n, \theta]$$

Trong xử lý tín hiệu số, g được định nghĩa như một filter-bank gồm các bộ lọc (filter) băng thông hình chữ nhật. Trong miền tần số, độ lớn của một bộ lọc băng thông tổng quát có thể được tính như hiệu số giữa 2 bộ lọc thông tần số thấp
Với f_1 f_2 lần lượt là tần số cắt thấp (low) và cao (high) đã được học, $rect(.)$ là hàm rectangular trong miền tần số.

$$G[f, f_1, f_2] = rect\left(\frac{f}{2f_2}\right) - rect\left(\frac{f}{2f_1}\right)$$

$$\xrightarrow{\text{Fourier Inverse}} g[n, f_1, f_2] = 2f_2 sinc(2\pi f_2 n) - 2f_1 sinc(2\pi f_1 n)$$

Công trình tiêu biểu: SincNet

Các kết quả đạt được

Speaker Identification Task

	TIMIT	LibriSpeech
DNN-MFCC	0.99	2.02
CNN-FBANK	0.86	1.55
CNN-Raw	1.65	1.00
SINCNET	**0.85**	**0.96**

Table 1: Classification Error Rate (CER%) of speaker identification systems trained on TIMIT (462 spks) and Librispeech (2484 spks) datasets. SincNets outperform the competing alternatives.

Speaker Verification Task

	d-vector	DNN-class
DNN-MFCC	0.88	0.72
CNN-FBANK	0.60	0.37
CNN-Raw	0.58	0.36
SINCNET	**0.51**	**0.32**

Table 2: Speaker Verification Equal Error Rate (EER%) on Librispeech datasets over different systems. SincNets outperform the competing alternatives.

Nhận xét

- Cơ sở lý thuyết Toán học vững vàng
- Tính toán nhanh và gọn nhẹ
- Kết hợp với Deep Learning một cách hiệu quả: Sử dùng DNN-Class trong đánh giá, cho kết quả đầy khả quan, có độ lỗi EER thấp

Thực nghiệm

Phần thực hành demo với SincNet


Kai Yu Nanxin Chen, Yanmin Qian.
Multi-task learning for text-dependent speaker verification.
In INTERSPEECH 2015 16th Annual Conference of the International Speech Communication Association, pages 185–189, 2015.


Mirco Ravanelli and Yoshua Bengio.
Speaker recognition from raw waveform with sincnet, 2019.


David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur.
X-vectors: Robust dnn embeddings for speaker recognition.
In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5329–5333, 2018.


Zhiyuan Tang, Lantian Li, and Dong Wang.
Multi-task recurrent model for speech and speaker recognition, 2016.


Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez.
Deep neural networks for small footprint text-dependent speaker verification.