

# MFCC and Prosodic Feature Extraction Techniques

Created by	Trí Trần
Created time	@May 2, 2024 1:40 AM

Link paper:

[https://www.researchgate.net/publication/258652444\\_MFCC\\_and\\_Prosodic\\_Feature\\_Extraction\\_Techniques\\_A\\_Compar](https://www.researchgate.net/publication/258652444_MFCC_and_Prosodic_Feature_Extraction_Techniques_A_Compar)

## Abstract

Trong bài báo này, mục tiêu chính của chúng tôi là cung cấp **sự khác biệt giữa các kỹ thuật trích xuất đặc trưng cepstral và không cepstral**. Ở đây, chúng tôi cố gắng đề cập đến hầu hết các **đặc trưng so sánh của hệ số cepstral tần số Mel và các đặc trưng prosodic**. Trong nhận dạng người nói, có hai loại kỹ thuật trích xuất đặc trưng có sẵn: **đặc trưng ngắn hạn tức là Hệ số Cepstral Tần số Mel (MFCC) và kỹ thuật trích xuất đặc trưng dài hạn (Prosodic)**.

Trong bài báo này, chúng tôi khám phá tính hữu ích của các đặc trưng prosodic cho việc phân loại âm tiết và MFCC cho việc trích xuất đặc trưng của tín hiệu nói, tiếp theo là so sánh giữa chúng. **Hệ số Cepstral Tần số Mel (MFCC) là một trong những kỹ thuật trích xuất đặc trưng quan trọng nhất, cần thiết trong nhiều loại ứng dụng nói**. Các đặc trưng MFCC được trích xuất từ các âm tiết của người nói trong các câu nói được phân đoạn trước. Hiện nay, các đặc trưng prosodic đang được sử dụng trong hầu hết các thuật toán nhận diện cảm xúc. Các đặc trưng prosodic tương đối đơn giản trong cấu trúc của họ và được biết đến với hiệu quả trong một số nhiệm vụ nhận dạng tiếng nói. Có nhiều cách tạo ra các đặc trưng dạng đường nét âm tiết prosodic mà gần đây đã được áp dụng để tăng cường hệ thống nhận diện người nói.

## 1. Introduction

Bài báo tập trung vào việc **so sánh giữa các đặc trưng phổ (như MFCC) và các đặc trưng prosodic trong hệ thống nhận dạng người nói**. Điều này nhấn mạnh rằng trong điều kiện âm thanh sạch hoặc tương tự, các đặc trưng phổ có thể hiệu quả, nhưng gặp khó khăn khi đối mặt với biến thể trong điều kiện thực tế. Do đó, việc mô hình hóa các đặc trưng có phạm vi dài như từ vựng, prosodic và thói quen diễn đạt trong nhận dạng người nói tự động trở nên quan trọng.

Các đặc trưng có phạm vi dài này không chỉ có thể cải thiện hiệu suất so với các đặc trưng phổ, mà còn có thể **phản ánh hành vi tự ý và đặc điểm của tiếng nói, chẳng hạn như phong cách nói. Sự khác biệt trong cấu trúc sinh lý của đường ống thoại của mỗi người được phản ánh trong phổ tần số của tín hiệu tiếng nói**, và điều này có thể được sử dụng cho việc nhận dạng người nói.

Hệ thống nhận dạng người nói tự động thường được chia thành hai giai đoạn chính: trích xuất đặc trưng và phân loại người nói. **MFCC là một đặc trưng phổ biến trong trích xuất đặc trưng**, trong khi các **đặc trưng prosodic cũng được sử dụng để tăng cường hệ thống nhận dạng người nói**.

Tóm lại, **việc nghiên cứu và áp dụng các đặc trưng có phạm vi dài như từ vựng, prosodic và thói quen diễn đạt trong hệ thống nhận dạng người nói tự động là rất quan trọng để cải thiện hiệu suất và độ chính xác của các ứng dụng nhận dạng người nói**.

## 2. MEL FREQUENCY CEPSTRAL COEFFICIENT

Đối với nhận dạng người nói, một kỹ thuật trích xuất đặc trưng cần phải trích xuất cả các đặc trưng tuyến tính và phi tuyến. Trong trường hợp này, chúng ta sử dụng các **Hệ số Cepstral Tần số Mel (MFCC)**. MFCC là một loại sóng đa phổ trong đó các tỷ lệ tần số được đặt trên một thang tần số tuyến tính cho các tần số nhỏ hơn 1 kHz và trên một thang log cho các tần số lớn hơn 1 kHz. **MFCC có khả năng bắt các đặc điểm quan trọng của tín hiệu âm thanh**.

Các hệ số cepstral phức tạp được gọi là MFCC. **MFCC chứa cả thông tin thời gian và tần số của tín hiệu, điều này làm cho chúng trở nên hữu ích hơn cho việc trích xuất đặc trưng**. MFCC đã được sử dụng rộng rãi trong lĩnh vực nhận dạng tiếng nói và đã xử lý được các đặc trưng động vì chúng trích xuất cả các thuộc tính tuyến tính và phi tuyến của

**tín hiệu.** MFCC có thể là một công cụ hữu ích cho việc trích xuất đặc trưng trong tín hiệu rung vì **rung động chứa cả các đặc trưng tuyến tính và phi tuyến.**

Như nhiều nghiên cứu đã chỉ ra, các đặc trưng phổ biến nhất được sử dụng trong các hệ thống xác minh/người nhận dạng người nói SOTA hiện đại là MFCC.



You

Đặc trưng tuyến tính và phi tuyến tính là gì



ChatGPT

Trong ngữ cảnh của xử lý tín hiệu, đặc trưng tuyến tính và phi tuyến tính là hai loại đặc trưng được sử dụng để mô tả các thuộc tính của tín hiệu.

**1. Đặc trưng tuyến tính:** Đặc trưng tuyến tính là các thuộc tính của tín hiệu mà có thể biểu diễn hoặc được phân tích một cách tuyến tính. Các ví dụ về đặc trưng tuyến tính bao gồm biên độ, tần số, và năng lượng. Khi tín hiệu được biểu diễn hoặc phân tích bằng các phương pháp tuyến tính như biến đổi Fourier, các đặc trưng này thường dễ dàng được trích xuất và sử dụng để mô tả tín hiệu.

**2. Đặc trưng phi tuyến tính:** Trái ngược với đặc trưng tuyến tính, các đặc trưng phi tuyến tính là các thuộc tính của tín hiệu mà không thể biểu diễn hoặc phân tích một cách tuyến tính. Các ví dụ về đặc trưng phi tuyến tính bao gồm các mối quan hệ không tuyến tính giữa các thành phần của tín hiệu, như các đặc điểm đường cong, sự biến đổi không đều theo thời gian, hoặc sự tương tác phức tạp giữa các thành phần. Đặc trưng phi tuyến tính thường đòi hỏi các phương pháp phức tạp hơn để trích xuất và hiểu, và thường cần được xử lý bằng các phương pháp không tuyến tính.

Trong ngữ cảnh của bài báo, việc sử dụng cả đặc trưng tuyến tính và phi tuyến tính (được biểu diễn thông qua MFCC) cho phép hệ thống nhận dạng người nói hiểu và mô hình hóa cả các thuộc tính đơn giản và phức tạp của tín hiệu âm thanh, từ đó cải thiện khả năng nhận dạng.

MFCC được sử dụng rộng rãi trong các hệ thống Nhận dạng Người Nói Tự Động vì

**1. Các đặc trưng cepstral gần như là độc lập với nhau vì DCT:**

- Khi tính toán các đặc trưng cepstral, chúng thường được chia thành các thành phần gọi là "cepstral coefficients" bằng cách sử dụng phép biến đổi cosin rời rạc (DCT).
- Sự sắp xếp này tạo ra sự độc lập tương đối giữa các đặc trưng cepstral, giúp chúng tách biệt và dễ dàng phân biệt các đặc điểm của tín hiệu âm thanh.

**2. Loại bỏ trung bình cepstral giúp loại bỏ tiếng ồn tĩnh từ kênh:**

- Bằng cách tính toán giá trị trung bình của các cepstral coefficients trong một tín hiệu âm thanh, chúng ta có thể loại bỏ tiếng ồn tĩnh gắn liền với kênh truyền dẫn.
- Quá trình này giúp làm sạch tín hiệu và tăng cường chất lượng của nó bằng cách giảm thiểu ảnh hưởng của tiếng ồn không mong muốn.

**3. MFCC ít nhạy cảm hơn với tiếng ồn thêm vào so với LPCC:**

- MFCC thường ít bị ảnh hưởng bởi tiếng ồn bổ sung hơn so với một số kỹ thuật trích xuất đặc trưng khác như LPCC.
- Điều này có nghĩa là MFCC có khả năng xử lý tốt hơn trong môi trường có tiếng ồn, giúp cải thiện độ chính xác của hệ thống nhận dạng người nói trong các điều kiện không thuận lợi.

Để trích xuất đặc trưng, MFCC thực hiện các bước sau:

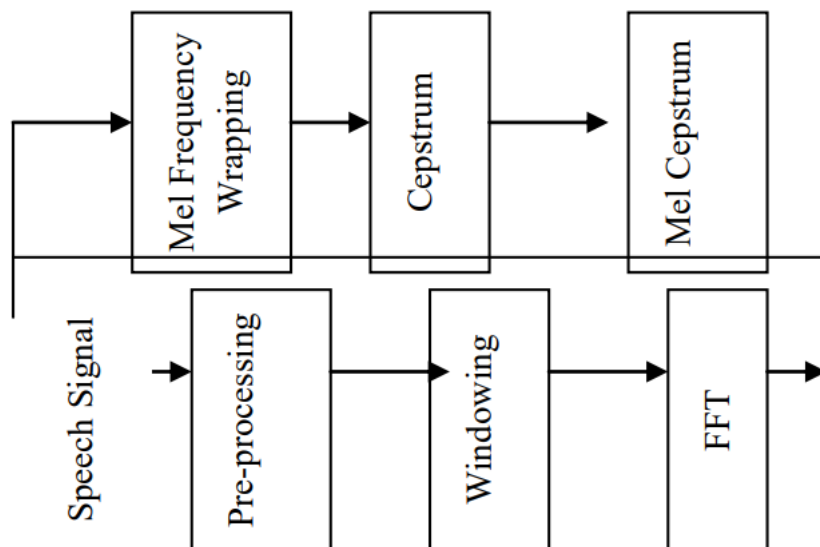
1. **Tiền xử lý tín hiệu:** Đầu tiên, một bước tiền xử lý được áp dụng trên tín hiệu nói. Điều này bao gồm việc sử dụng bộ lọc tiền đặt để làm phẳng biên độ tín hiệu và làm giảm kích thước chính xác của nó.
2. **Áp dụng cửa sổ Hamming:** Một cửa sổ Hamming được áp dụng lên mỗi khối để giảm thiểu hiệu ứng biên do việc cắt cửa sổ.
3. **Biến đổi Fourier nhanh (FFT):** Một biến đổi Fourier nhanh được áp dụng lên tín hiệu đã được xử lý để chuyển đổi nó từ miền thời gian sang miền tần số.
4. **Lọc bằng các bộ lọc tam giác trên thang Mel:** Tín hiệu sau khi được chuyển đổi Fourier được lọc thông qua một loạt các bộ lọc tam giác được phân bố trên một thang tần số Mel. Các bộ lọc này giúp tập trung vào các thành phần quan trọng của tín hiệu ở các dải tần số khác nhau.
5. **Tính toán các hệ số MFCC:** Sau khi tín hiệu đã được lọc bởi các bộ lọc tam giác Mel, các hệ số MFCC được tính toán từ tín hiệu đã được lọc này.

Bước cuối cùng trong quá trình tính toán của MFCC là việc tính toán thang Mel, thường được thực hiện bằng cách sử dụng các công thức phù hợp để phân bố các bộ lọc tam giác trên thang tần số Mel:

$$M = \frac{1000}{\log 2} \log \left( 1 + \frac{f}{1000} \right) \dots \dots \dots (1)$$

Where f is the frequency.

MFCC have the following steps when feature extracted from a speech signal-



**Fig. 1: MFCC steps for feature extraction**

**Bộ lọc** là thành phần chính của MFCC và đóng vai trò quan trọng trong việc tăng cường khả năng chống lại tiếng ồn. Cụ thể, các bộ lọc giúp làm mịn phổ tần số, làm giảm biến động do tiếng ồn thêm vào trên toàn bộ băng thông của mỗi bộ lọc. Điều này có nghĩa là các bộ lọc giúp loại bỏ hoặc giảm thiểu ảnh hưởng của tiếng ồn thêm vào, giúp tăng cường sự ổn định của đặc trưng MFCC và cải thiện khả năng nhận dạng người nói trong môi trường có nhiều tiếng ồn.



You

bộ lọc hoạt động như thế nào



ChatGPT

Bộ lọc trong MFCC hoạt động bằng cách chia tín hiệu âm thanh thành các dải tần số và làm mịn phổ tần số của mỗi dải này. Quá trình hoạt động của bộ lọc có thể được mô tả như sau:

1. **Chia tín hiệu thành các dải tần số:** Trước tiên, tín hiệu âm thanh được chia thành các dải tần số, mỗi dải tương ứng với một bộ lọc. Các dải tần số này thường được phân bố dựa trên thang tần số Mel để phản ánh cách thức người nói nghe và phản ứng với các tần số âm thanh.
2. **Làm mịn phổ tần số:** Mỗi bộ lọc thực hiện một phép làm mịn trên phổ tần số của tín hiệu âm thanh trong dải tần số tương ứng. Điều này có thể bao gồm việc áp dụng các hàm trọng số như cửa sổ tam giác hoặc hàm hanning cho mỗi phần của phổ tần số để giảm thiểu hiệu ứng biên.
3. **Giảm biến động do tiếng ồn:** Bằng cách làm mịn phổ tần số, các bộ lọc giúp giảm biến động do tiếng ồn thêm vào trên toàn bộ băng thông của mỗi bộ lọc. Điều này làm cho các đặc trưng được trích xuất từ mỗi dải tần số ít bị ảnh hưởng bởi tiếng ồn và ổn định hơn.

Tóm lại, bộ lọc trong MFCC chia tín hiệu thành các dải tần số và giúp làm mịn phổ tần số của mỗi dải, giảm biến động do tiếng ồn và cải thiện khả năng chống lại tiếng ồn của đặc trưng MFCC.

### 3. PROSODIC FEATURE EXTRACTION

Trong PROSODIC, có ba đặc tính chính là **cao độ**, **thời lượng** và **cường độ** được lấy từ bất kỳ tín hiệu nói nào để nhận dạng người nói. Như đã thảo luận trong nhiều bài báo, **Prosodic speech features** đã được biết đến để cung cấp thông tin hữu ích về phong cách nói của một người, và do đó, rộng rãi được sử dụng trong các ứng dụng nhận dạng người nói. Một đặc điểm ngữ điệu được trích xuất ở mức từ ngữ được sử dụng cho các nhiệm vụ khác nhau, như: **phát hiện căng thẳng tự động**, **nhận dạng người nói** và thậm chí là **mô hình hóa ngôn ngữ**. Các đặc điểm cao độ quan trọng nhất là những đặc điểm **nắm bắt mức độ cao độ**, trong khi các đặc điểm **năng lượng quan trọng nhất phản ánh các mẫu tăng và giảm mức độ năng lượng**. Đối với các đặc điểm thời lượng, thời lượng của hạt nhân (nucleus) quan trọng hơn đối với việc nhận dạng người nói so với thời lượng từ bắt đầu của một từ ngữ. Các đặc điểm ngữ điệu chứa đựng các đặc tính dài hạn vì chúng cung cấp một mô tả về các thuộc tính thường gặp của người nói. Đặc tính chính là cao độ và năng lượng có hiệu suất mạnh mẽ trong việc nhận dạng người nói đặc biệt khi dữ liệu không đồng nhất và kênh không phù hợp. Ngoài ra, các đặc điểm ngữ điệu chứa thông tin cụ thể về người nói, do sự khác biệt về cấu trúc dây thanh âm giữa các người nói. **Mặt không thực tế của các đặc điểm ngữ điệu là lượng dữ liệu lớn cần thiết cho việc nhận dạng thành công**, cũng như quy trình cần thiết để thu được chúng là **phức tạp và tốn nhiều tài nguyên tính toán** vì các đặc điểm ngữ điệu được cho là được chuyển đạt bởi các từ ngữ trong việc phân đoạn tiếng nói được thực hiện trước tiên để thu được các đơn vị giống như từ ngữ gọi là pseudo syllables.

- Các đặc điểm **cepstral ngắn hạn** thường được gọi là các **đặc điểm cấp thấp**, tập trung vào các **thông số giọng của người nói**, trong khi các **đặc điểm cấp cao hơn** như **ngữ âm**, **ngữ điệu** và **từ vựng**.
- Các **prosodic ghi lại sự biến đổi trong cách nói, thời gian và âm lượng cụ thể cho từng người nói**. Chúng là các đặc điểm siêu phân đoạn, mở rộng ra ngoài một phần, và thường được coi là một phần của các đặc điểm dài hạn.
- Có nhiều thách thức khi tính toán các **prosodic**, bao gồm **xác định phần của tín hiệu nói có ích để tìm thông tin**, **chọn mô hình tính toán phù hợp để cải thiện hiệu suất của prosodic**, và **đánh giá tính đồng bộ và hiệu quả khi sử dụng đơn lẻ hoặc kết hợp với các đặc trưng khác như đặc điểm cepstral**.

TR

You

Vậy tóm lại Prosodic lấy những thông tin gì để đưa ra dự đoán

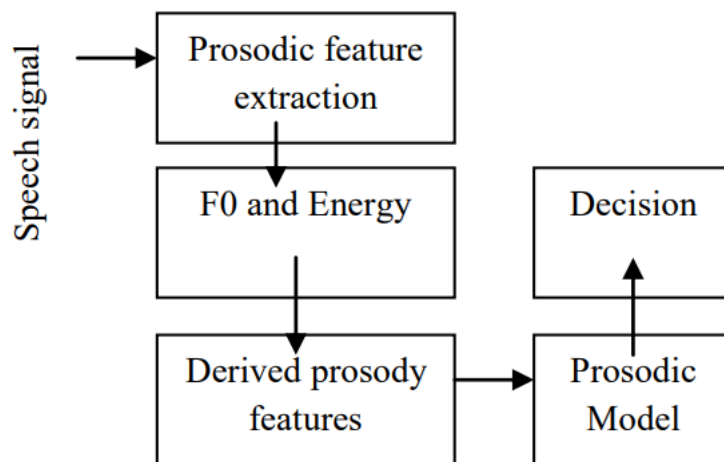
TR

ChatGPT

Prosodic lấy các thông tin về cao độ (pitch), thời lượng (duration), và cường độ (loudness) từ tín hiệu nói để đưa ra dự đoán. Các đặc điểm này cho phép nắm bắt các biến thể trong cách nói, âm điệu và nhịp điệu mà là đặc trưng cụ thể cho từng người nói. Bằng cách phân tích các biến thể này, prosodic có thể giúp nhận biết các đặc điểm độc nhất của người nói và cung cấp thông tin quan trọng để nhận dạng hoặc phân loại người nói.

🗨️ 🗨️ 🗨️ 🗨️

Những **prosodic** dựa trên **cao độ** có thể **ít bị ảnh hưởng** hơn bởi **hiệu ứng của thiết bị và kênh truyền**. Quá trình trích xuất đặc trưng ngữ điệu từ tín hiệu nói bao gồm các bước sau:



**Fig 2: Steps of prosodic feature extraction**

**Mô hình trích xuất đặc trưng** là một mô hình xác suất, sử dụng **cây quyết định** để tính **tỷ lệ nhận dạng**. Sau khi trích xuất đặc trưng, một khi kết quả nhận dạng với các phân đoạn chi tiết thời gian được có sẵn, chúng ta có thể bắt đầu **mô hình hóa các đặc trưng vượt ra ngoài các âm và từ**. Một khía cạnh quan trọng của biến thể ngữ điệu là thời lượng của các đơn vị.

TR

You

Mô hình trích xuất đặc trưng là một mô hình xác suất, là gì

TR

ChatGPT

Mô hình trích xuất đặc trưng là một mô hình xác suất được sử dụng để xác định các đặc trưng quan trọng từ dữ liệu. Mô hình này tính toán xác suất của các đặc trưng cụ thể xuất hiện trong dữ liệu và sử dụng thông tin này để quyết định xem một đặc trưng nào nên được trích xuất và đưa vào quá trình nhận dạng hoặc phân loại. Đối với mô hình trích xuất đặc trưng dựa trên cây quyết định, ví dụ, các quy tắc quyết định được áp dụng dựa trên xác suất xuất hiện của các đặc trưng trong dữ liệu huấn luyện để xác định cách trích xuất đặc trưng hiệu quả nhất.

🗨️ 🗨️ 🗨️ 🗨️

## 4. A COMPARATIVE STUDY ON MFCC AND PROSODIC


1. **Các đặc điểm tiếng nói được sử dụng:** Trong hệ thống nhận dạng người nói, sử dụng cả các đặc điểm dựa trên phổ và các đặc điểm ngữ điệu để xác định và phân loại người nói.
2. **Trọng tâm của việc trích xuất đặc trưng:** Việc trích xuất và lựa chọn biểu diễn thông số tối ưu của tín hiệu âm thanh là quan trọng và ảnh hưởng đến hiệu suất nhận dạng người nói.
3. **So sánh giữa MFCC và Prosodic:** Có nhiều nghiên cứu đã thực hiện về cả hai loại đặc điểm này, cho thấy rằng MFCC thường sử dụng một tập hợp nhỏ các đặc điểm tiêu chuẩn, trong khi Prosodic sử dụng các đặc điểm dài hạn.
4. **Động lực của việc sử dụng các đặc điểm dài hạn:** Trong lĩnh vực nhận dạng người nói, đã có các nghiên cứu khác nhau để sử dụng thông tin động có trong tiếng nói, nhưng các đặc điểm dài hạn đòi hỏi nhiều mẫu tiếng nói và làm phức tạp quá trình tính toán.
5. **Ý nghĩa của đặc điểm dài hạn trong nhận dạng người nói:** Các đặc điểm dài hạn có thể cung cấp thông tin quan trọng hơn cho việc phân biệt người nói và có thể phát hiện ra các đặc điểm cá nhân của giọng nói và hành vi nói của họ một cách chi tiết.

Giới hạn chính của MFCC:

1. **Sự phụ thuộc vào thiết kế của bộ lọc:** Một hạn chế nghiêm trọng của kỹ thuật trích xuất đặc trưng MFCC ban đầu là băng thông của bộ lọc không phải là một tham số thiết kế độc lập mà thay vào đó được xác định bởi phạm vi tần số của bộ lọc và số lượng bộ lọc được sử dụng trong quá trình nhận dạng. Do đó, số lượng bộ lọc có thể được thêm hoặc giảm để cung cấp tần số lấy mẫu của các thí nghiệm.
2. **Ảnh hưởng của số lượng bộ lọc và loại cửa sổ:** Hiệu suất của các hệ số MFCC có thể bị ảnh hưởng bởi số lượng bộ lọc và loại cửa sổ được sử dụng. Chính xác của kết quả có thể biến động khi tăng hoặc giảm số lượng hệ số MFCC. Sự sử dụng quá ít hoặc quá nhiều bộ lọc cũng không dẫn đến độ chính xác tốt hơn. Ngoài ra, hiệu quả cao nhất được mô tả khi sử dụng cửa sổ Hanning.

 You

Nghĩa là MFCC phụ thuộc quá lớn vào cửa sổ lọc đúng không

 ChatGPT

Đúng, MFCC phụ thuộc mạnh mẽ vào việc chọn loại cửa sổ lọc. Việc chọn cửa sổ lọc không phù hợp có thể ảnh hưởng đến hiệu suất của quá trình trích xuất đặc trưng MFCC. Trong nghiên cứu, đã chỉ ra rằng việc sử dụng cửa sổ lọc Hanning mang lại hiệu suất cao nhất cho quá trình này.

Tóm lại, theo các nghiên cứu, **prosodic** và **các tham số liên quan đến cách nói của một người dễ dàng bị mô phỏng hơn so với các tham số mô tả cấu trúc của ống hô hấp**. Do đó, trong việc áp dụng công nghệ xác minh người nói trong thực tế, việc sử dụng các tham số ngữ điệu cần phải được thực hiện cẩn thận, đặc biệt là trong các ứng dụng mà có rủi ro cao về việc cố gắng gian lận hoặc trong những trường hợp mà việc chấp nhận nhầm người mạo danh có thể gây tổn thất nghiêm trọng.

**You**

Vậy prosodic dễ bị bắt chước hơn so với các đặc tính thanh âm của người nói đúng không

**ChatGPT**

Đúng, đó là điều mà các nghiên cứu đã chỉ ra. Các đặc điểm ngữ điệu, như cách người nói nhấn mạnh, điều chỉnh âm lượng và tốc độ, thường dễ bị bắt chước hơn so với các đặc điểm về cấu trúc của âm học của giọng nói. Điều này khiến cho việc sử dụng các tham số ngữ điệu trong các hệ thống xác minh người nói đòi hỏi sự cẩn trọng đặc biệt để ngăn chặn việc giả mạo hoặc lừa đảo.

🔊 📄 ↺ 🔍

**You**

Cách nói,... là các đặc điểm phi tuyến tính đúng không

**ChatGPT**

Đúng, cách nói, bao gồm cách nhấn mạnh, tốc độ và điều chỉnh âm lượng, là các đặc điểm phi tuyến tính của giọng nói. Điều này có nghĩa là chúng không phải là các đặc điểm tuyến tính và thường không thể mô tả bằng các quy luật hoặc hàm số đơn giản.

🔊 📄 ↺ 🔍

Dựa trên nghiên cứu so sánh từ nhiều bài báo, các hệ thống sử dụng sự kết hợp giữa các đặc điểm cepstral và các đặc điểm cấp cao ("đặc điểm cấp cao" thường đề cập đến các đặc điểm ngữ âm không phụ thuộc vào các đặc điểm cụ thể của giọng nói hoặc ngữ âm, mà thay vào đó liên quan đến các yếu tố như cú pháp ngôn ngữ, mẫu thời gian, hoặc cách sử dụng từ vựng. Các đặc điểm cấp cao có thể bao gồm cả thông tin ngữ âm dài hạn như ngữ điệu, nhấn mạnh, và tốc độ, cũng như thông tin về ngữ cảnh và cách sử dụng ngôn ngữ. Trong bối cảnh này, các đặc điểm cấp cao thường được sử dụng để bổ sung và tối ưu hóa hiệu suất của các hệ thống xác minh người nói.) thường có hiệu suất vượt trội so với các hệ thống tiêu chuẩn, đặc biệt là khi có nhiều dữ liệu huấn luyện. Các hệ thống cepstral ngắn hạn thường thực hiện tốt bởi vì chúng phản ánh thông tin về sinh lý của người nói và không phụ thuộc vào nội dung ngữ âm. Tuy nhiên, thông tin dài hạn trong tín hiệu cũng cần được sử dụng để tối ưu hóa hiệu suất. Các đặc điểm cấp cao cũng có thể cung cấp độ bền cao hơn đối với biến đổi kênh, vì chúng không phụ thuộc vào điều kiện âm thanh cụ thể.

Tóm lại, các đặc điểm dài hạn đã được nghiên cứu trong nhiều năm và có dấu hiệu cho thấy chúng có thể hữu ích cho việc nhận diện người nói. Hiện nay, các hệ thống xác minh người nói sử dụng các loại đặc điểm ngữ điệu khác nhau đã được đề xuất. Mặc dù đã được chỉ ra rằng hầu hết các hệ thống xác minh người nói này có thể cải thiện hiệu suất hệ thống bằng cách kết hợp điểm số với các hệ thống dựa trên đặc điểm cepstral hiện đại, nhưng việc so sánh hệ thống mô hình hóa ngữ điệu được sử dụng trong những hệ thống ngữ điệu này một cách có hệ thống vẫn chưa được thực hiện.

Tóm lại, các đặc điểm ngữ điệu như tần số cơ bản và biên độ của giọng nói cung cấp thông tin quan trọng cho việc nhận diện/xác minh người nói. Mặc dù đã từng được nghiên cứu, sự quan tâm vào việc sử dụng các đặc điểm ngữ điệu giảm đi do chúng không đem lại hiệu suất mong muốn, nhưng chúng có thể được sử dụng để cải thiện hiệu suất của các thuật toán sử dụng Mô hình Markov Ẩn, đặc biệt trong bối cảnh của các vấn đề như biến dạng kênh và tiếng ồn.

Tóm lại, dựa trên các nghiên cứu, độ chính xác trung bình của tập hợp đặc điểm MFCC là trên 95%, trong khi đó, sử dụng tập hợp prosodic của tín hiệu giọng nói có tỷ lệ chính xác trung bình là 69%. Có nhiều nghiên cứu cho thấy rằng phân tích dựa trên hình thức ngữ âm dựa trên MFCC, thể hiện mô hình tại, đã cho kết quả tốt trong việc nhận diện người nói chủ yếu khi sử dụng một số lượng cao các hệ số. Ngoài ra, nó được coi là hệ thống nhận diện người nói



thành công nhất khi đối mặt với các biến thể khác nhau như: ngữ điệu, ngữ điệu, tiếng ồn, v.v. Nó cũng thực hiện các nhiệm vụ lọc, mô hình hóa và xử lý, giải mã, âm vị hoặc từ và phân biệt ngôn ngữ. Bảng thể hiện rằng một số tham số phân biệt MFCC và ngữ điệu sao cho ở một số nơi đặc điểm cepstral (đặc điểm ngắn hạn) tốt hơn và ở một số nơi đặc điểm dài hạn hoặc ngữ điệu mang lại kết quả tốt hơn. Chúng tôi đã thảo luận một số tham số của MFCC và ngữ điệu trong bảng dưới đây.

**Table 1. MFCC & Prosodic: A comparative Chart**

Parameter	MFCC	Prosodic
Vocal Tract	Depends on shape of the vocal tract	Excitation of the vocal tract and the speaking style
Features set	Uses a small set of standard features	Uses long term features
Features type	Uses Cepstral features	Uses Non cepstral features
Secure	Not easy to mimic	easier to mimic
Filters	Uses filter bank	Does not use filters
Speaker Verification/Identification	It give better results for both i.e. SI/ SV	It give better result for SV
Channel effect	Cepstral features affected by the channel distortion.	It is believed that prosodic features are less vulnerable to the channel distortion.
Performance	MFCC lonely able to perform well.	Prosodic features alone cannot perform well.
Speech Sample	It requires less Speech sample less time and not so computationally complex.	It require a lot of speech samples and are also time consuming and computationally complex

## 5. SUMMARY

Tìm hiểu về Nhận dạng Người nói, đã được tìm thấy rằng có nhiều yếu tố ảnh hưởng đến hiệu suất của quá trình nhận dạng người nói như biến thể giọng nói, không phù hợp với kênh, **các loại điện thoại/micro khác nhau**, v.v. Trong những yếu tố này, **tiếng ồn có tác động mạnh mẽ nhất đến hiệu suất của quá trình nhận dạng người nói**. Nghiên cứu cũng chỉ ra rằng có một số tham số khác ảnh hưởng đến hiệu suất của quá trình nhận dạng người nói, bao gồm:

- **Ngôn ngữ được sử dụng trong dữ liệu huấn luyện và kiểm tra.**
- **Chất lượng của dữ liệu mẫu giọng nói trong quá trình huấn luyện và kiểm tra.**
- **Chất lượng của microphone và khoảng cách giữa microphone và người nói.**
- **Tiếng ồn trong quá trình ghi âm mẫu giọng nói và kiểm tra giọng nói.**



- Độ dài của mẫu giọng nói được sử dụng trong quá trình huấn luyện và kiểm tra.
- Bảo vệ microphone hoặc loa vào thời gian huấn luyện và kiểm tra.
- Phụ thuộc vào văn bản, tức là dữ liệu huấn luyện và kiểm tra là giống nhau.
- Sự biến đổi trong giọng nói của người nói.

Hiệu suất của quá trình nhận dạng người nói bị ảnh hưởng bởi các yếu tố được đề cập ở trên và một số yếu tố khác nữa. Sự biến đổi trong giọng nói của người nói cũng là yếu tố chính ảnh hưởng đến quá trình nhận dạng người nói.

## 6. Conclusion

Trong bài báo này, chúng tôi cố gắng bao quát **những khác biệt cơ bản giữa MFCC và Prosodic**. Công việc này phản ánh các kết quả thu được trong việc đánh giá các đặc điểm của các đặc trưng Prosodic và MFCC. Trong nghiên cứu này, chúng tôi trình bày các kỹ thuật trích xuất đặc trưng cho việc nhận dạng người nói, trong đó đã thảo luận về MFCC và Prosodic. Kết luận là **dựa trên quan điểm của các nhà nghiên cứu và thực thi, MFCC tốt hơn so với Prosodic và được biết đến là các kỹ thuật phổ biến được sử dụng trong nhận dạng người nói để mô tả các đặc điểm của tín hiệu, liên quan đến các thuộc tính của dải phát âm có khả năng phân biệt người nói.**

Ngoài ra, khái niệm được đưa ra là **các hệ thống sử dụng kết hợp của các đặc trưng cepstral và cấp cao vượt trội hơn so với các hệ thống tiêu chuẩn, đặc biệt là khi lượng dữ liệu huấn luyện có sẵn tăng lên.** Điều này xác nhận giả định rằng các hệ thống cepstral ngắn hạn thường thực hiện tốt vì chúng phản ánh thông tin về cấu trúc vật lý của người nói và không phụ thuộc vào nội dung phát âm. Tuy nhiên, **thông tin về phạm vi dài hạn cũng tồn tại trong tín hiệu chỉ được khai thác trong các hệ thống kết hợp.** Như nghiên cứu đã chỉ ra rằng **MFCC là tốt nhất nhưng các đặc trưng cấp cao cũng có tiềm năng tăng sự ổn định đối với biến thể kênh, vì việc sử dụng từ vựng hoặc mẫu thời gian.**