**Report**

**Audio Analysis Techniques**

*Trong quá trình kiểm thử chương trình, nhiều trường hợp nhận diện sai thành viên Lê Anh Tuấn (với vai trò là bố mẹ). Do đó cần tìm cách cải thiện mô hình. Bài viết này đề cập đến việc tìm hiểu các kỹ thuật phân tích âm thanh cho nhân dạng giọng nói. Cũng như trong tương lai nếu cần thiết thay đổi mô hình Nhận diện giọng nói (Speaker Recognition hay còn được gọi là Voice Recognition) có thể áp dụng để cải thiện độ chính xác.*

1. **Linear Predictive Coding(LPC):**

Linear Predictive Coding (LPC) là một kỹ thuật trong xử lý tín hiệu âm thanh để mô hình hóa và biểu diễn các thông tin trong tín hiệu âm thanh. LPC được sử dụng rộng rãi trong các ứng dụng như nén âm thanh, nhận dạng tiếng nói, và tổng hợp tiếng nói.

Cơ bản, LPC xác định một mô hình tuyến tính của tín hiệu âm thanh bằng cách dự đoán các giá trị tương lai của tín hiệu dựa trên các giá trị hiện tại và trước đó. Quá trình này thường bao gồm các bước sau:

* **Framing:** Tín hiệu âm thanh được chia thành các khung nhỏ, mỗi khung thường có độ dài cố định. Việc chia khung giúp làm giảm biến động và giúp LPC dễ dàng dự đoán các giá trị tương lai của tín hiệu.
* **Windowing:** Các khung tín hiệu được áp dụng hàm cửa (window function) để giảm thiểu hiện tượng bị méo, cũng như giảm thiểu hiện tượng bị méo bởi biên của các khung.
* **Pre-emphasis:** Tín hiệu âm thanh thường bị tắc nghẽn ở các tần số thấp. Để cải thiện tỷ lệ tín hiệu/nhiễu và giảm bớt hiện tượng tắc nghẽn, một bước pre-emphasis được áp dụng để tăng cường các tần số cao hơn.
* **Autocorrelation:** Tính toán hệ số tự tương quan của tín hiệu trong mỗi khung.
* **Levinson-Durbin recursion:** Sử dụng thuật toán Levinson-Durbin để ước lượng các hệ số của mô hình LPC từ các hệ số tự tương quan.
* **Residual computation:** Tính toán tín hiệu residual bằng cách trừ tín hiệu dự đoán từ tín hiệu ban đầu.

Kết quả của quá trình LPC là một tập hợp các hệ số, thường được gọi là vectơ LPC, mô tả các đặc điểm của tín hiệu âm thanh.

1. **Pure Fast Fourier Transform(Pure FFT):**

Pure FFT (Fast Fourier Transform) là một phương pháp để thực hiện biến đổi Fourier trên tín hiệu số sử dụng thuật toán nhanh Fourier. Trong ngữ cảnh của xử lý tín hiệu, FFT thường được sử dụng để chuyển đổi một tín hiệu từ miền thời gian sang miền tần số, cho phép phân tích và xử lý tín hiệu theo các thành phần tần số khác nhau.

Phương pháp FFT giúp giảm thiểu đáng kể thời gian tính toán so với biến đổi Fourier truyền thống, đặc biệt là khi số mẫu của tín hiệu là một lũy thừa của 2 (như 1024, 2048, 4096,...). Điều này làm cho FFT trở thành một công cụ quan trọng và phổ biến trong xử lý tín hiệu số.

Quy trình của FFT bao gồm các bước sau:

* **Chuẩn bị dữ liệu đầu vào:** Dữ liệu đầu vào (thường là tín hiệu âm thanh) được chia thành các khung có độ dài cố định. Điều này đảm bảo rằng độ dài của mỗi khung phải là một lũy thừa của 2 để áp dụng FFT hiệu quả.
* **Áp dụng hàm cửa (windowing):** Mỗi khung dữ liệu thường được nhân với một hàm cửa như cửa Hanning hoặc cửa Hamming để giảm thiểu hiện tượng bị méo.
* **Thực hiện FFT:** FFT được áp dụng cho mỗi khung dữ liệu để chuyển đổi từ miền thời gian sang miền tần số.
* **Tính toán biên độ và pha:** Biên độ và pha của các thành phần tần số được tính toán từ đầu ra của FFT.
* **Tính toán năng lượng:** Năng lượng của mỗi thành phần tần số được tính bằng cách lấy bình phương của biên độ.

1. **Power Spectral Analysis:**

Power Spectral Analysis là một kỹ thuật trong xử lý tín hiệu để phân tích năng lượng của tín hiệu trong miền tần số. Nó là quá trình phân tích tín hiệu để xác định cường độ của các thành phần tần số khác nhau trong tín hiệu, thường được biểu diễn dưới dạng biểu đồ Power Spectral Density (PSD) hoặc biểu đồ Spectrogram.

Các bước cơ bản của phân tích phổ bao gồm:

* **Chuẩn bị tín hiệu đầu vào:** Tín hiệu âm thanh thường được chia thành các khung nhỏ với kích thước cố định để phù hợp với quá trình phân tích phổ.
* **Áp dụng hàm cửa (windowing):** Mỗi khung dữ liệu thường được nhân với một hàm cửa như cửa Hanning hoặc cửa Hamming để giảm thiểu hiện tượng bị méo.
* **Thực hiện biến đổi Fourier (FFT hoặc DFT):** Biến đổi Fourier được áp dụng cho mỗi khung dữ liệu để chuyển đổi từ miền thời gian sang miền tần số. Kết quả là một phổ tần số của tín hiệu, biểu thị cường độ của các thành phần tần số khác nhau.
* **Tính toán Power Spectral Density (PSD):** Cường độ của mỗi thành phần tần số được tính toán bằng cách lấy bình phương của biên độ tương ứng.
* **Biểu diễn kết quả:** Kết quả PSD thường được biểu diễn dưới dạng biểu đồ, với trục tung là cường độ hoặc năng lượng và trục hoành là tần số.

Power Spectral Analysis *giúp hiểu rõ hơn về cấu trúc tần số của tín hiệu* và có thể được sử dụng để trích xuất đặc trưng, phát hiện sự kiện hoặc biến đổi tín hiệu.

1. **Perceptual Linear Prediction(PLP):**

Perceptual Linear Prediction (PLP) là một phương pháp trong xử lý tiếng nói để trích xuất đặc trưng từ tín hiệu âm thanh dựa trên các tính chất lẫn nhau của tai người và mô hình âm thanh như cơ chế sinh học và cơ chế tương tự máy.

Cụ thể, PLP kết hợp giữa Linear Prediction Cepstral Coefficients (LPCC) và Mel-Frequency Cepstral Coefficients (MFCC) để tạo ra các hệ số có tính biểu diễn tốt và chứa thông tin hữu ích về tín hiệu âm thanh. Quá trình PLP thường bao gồm các bước sau:

* **Framing và Windowing:** Tín hiệu âm thanh được chia thành các khung nhỏ có độ dài cố định và được áp dụng hàm cửa để giảm thiểu hiện tượng bị méo.
* **Pre-emphasis:** Một bước pre-emphasis có thể được áp dụng để tăng cường các tần số cao của tín hiệu và giảm nhiễu.
* **Linear Prediction (LP) Analysis:** Áp dụng phân tích Linear Prediction để ước lượng các hệ số dự đoán tuyến tính của tín hiệu.
* **Conversion to Mel Scale:** Chuyển đổi các tần số từ tần số Hertz sang tần số Mel sử dụng công thức đối với MFCC.
* **Discrete Cosine Transform (DCT):** Áp dụng biến đổi cosine rời rạc để chuyển đổi hệ số Mel sang miền cepstral.
* **Cepstral Mean Normalization (CMN):** Thực hiện việc chuẩn hóa cepstral bằng cách trừ đi trung bình cepstral của tất cả các khung.

Kết quả của quá trình PLP là một vectơ đặc trưng có chứa các hệ số cepstral có tính biểu diễn tốt cho tín hiệu âm thanh. Vectơ này thường được sử dụng trong các ứng dụng nhận dạng tiếng nói, nhận dạng người nói, và xử lý tiếng nói khác. PLP thường được ưa chuộng vì khả năng biểu diễn tốt của nó và sự tương tự với cách tai người nhận dạng và xử lý âm thanh.

1. **MEL Scale Cepstral Analysis(MEL):**

MEL Scale Cepstral Analysis (MEL) là một phương pháp trong xử lý tín hiệu âm thanh để trích xuất đặc trưng từ tín hiệu âm thanh dựa trên các đặc điểm về cảm giác của tai người trong việc nghe và nhận dạng âm thanh.

Cụ thể, MEL Scale Cepstral Analysis thường bao gồm các bước sau:

* **Framing và Windowing:** Tín hiệu âm thanh được chia thành các khung nhỏ với độ dài cố định và áp dụng hàm cửa để giảm thiểu hiện tượng bị méo.
* **Discrete Fourier Transform (DFT):** Áp dụng biến đổi Fourier rời rạc (DFT) cho mỗi khung dữ liệu để chuyển đổi từ miền thời gian sang miền tần số.
* **Filter Banks:** Tạo ra các bộ lọc Mel (hay còn gọi là Mel filter banks) theo thang Mel scale. Filter banks này được thiết kế để mô phỏng cách tai người nhận dạng các tần số âm thanh.
* **Filtering và Logarithmic Compression:** Áp dụng các bộ lọc Mel vào biến đổi Fourier của từng khung và sau đó thực hiện nén logarithmic trên kết quả để tăng cường sự phân biệt giữa các băng thông tần số.
* **Discrete Cosine Transform (DCT):** Áp dụng biến đổi cosine rời rạc (DCT) cho từng khung dữ liệu sau khi nén logarithmic để chuyển đổi từ miền tần số sang miền cepstral.
* **Cepstral Mean Normalization (CMN):** Thực hiện việc chuẩn hóa cepstral bằng cách trừ đi trung bình cepstral của tất cả các khung.

1. **Mel-frequency cepstral coefficients(MFCCs):**

Mel-frequency cepstral coefficients (MFCCs) là một phương pháp phổ biến để trích xuất đặc trưng từ tín hiệu âm thanh, thường được sử dụng trong các ứng dụng nhận dạng tiếng nói, nhận dạng người nói, và xử lý ngôn ngữ tự nhiên.

Ở mức cơ bản, MFCCs là các hệ số cepstral được tính toán từ biến đổi Fourier của tín hiệu âm thanh, với mục tiêu là biểu diễn các đặc trưng phân cấp của tín hiệu âm thanh dựa trên cảm giác nghe của con người.

Quá trình tính toán MFCCs bao gồm các bước chính sau:

* **Pre-emphasis:** Áp dụng bộ lọc pre-emphasis để làm tôn lên các thành phần cao tần của tín hiệu âm thanh.
* **Framing:** Chia tín hiệu âm thanh thành các khung (frame) nhỏ có kích thước cố định. Mỗi frame thường có độ dài khoảng vài mili giây.
* **Hamming Windowing:** Áp dụng cửa có hàm Hamming cho từng frame để giảm thiểu hiện tượng "leakage" khi thực hiện biến đổi Fourier.
* **FFT (Fast Fourier Transform):** Thực hiện biến đổi Fourier nhanh (FFT) trên mỗi frame để chuyển đổi tín hiệu âm thanh từ miền thời gian sang miền tần số.
* **Mel Filterbank:** Tính toán năng lượng trong mỗi frame qua một bộ lọc Mel để tạo ra một biểu diễn tần số dựa trên đặc điểm của cấu trúc cảm giác nghe của con người.
* **Logarithm:** Lấy logarit của năng lượng tính được từ bước trước, để tương thích với cách con người cảm nhận âm thanh.
* **Discrete Cosine Transform (DCT):** Thực hiện biến đổi cosin rời rạc (DCT) trên các hệ số cepstral của mỗi frame để thu được MFCCs.

Cuối cùng, MFCCs được biểu diễn dưới dạng ma trận, trong đó mỗi hàng tương ứng với một hệ số cepstral và mỗi cột tương ứng với một frame của tín hiệu âm thanh.

1. **Human Factor Cepstral Coefficients(HFCC):**

HFCC là viết tắt của "Human Factor Cepstral Coefficients", một phương pháp trích xuất đặc trưng âm thanh trong lĩnh vực xử lý tín hiệu âm thanh. Đây là một phương pháp được thiết kế để cải thiện hiệu suất của các hệ thống nhận dạng tiếng nói trong môi trường có nhiều nhiễu và méo tiếng, bằng cách tích hợp các yếu tố của hệ thống thính giác của con người vào quá trình trích xuất đặc trưng.

Dưới đây là các bước thực hiện của HFCC:

* **Tính toán Mel-frequency filter bank:** Đầu tiên, tạo ra một bộ lọc với các băng thông tuyến tính trên thang tần số Mel. Các bộ lọc này thường được phân bố không đều trên thang tần số để phù hợp với cách con người nghe.
* **Tính toán log-energy spectrum:** Sử dụng DFT (Discrete Fourier Transform), tính toán phổ năng lượng logarithmic của tín hiệu âm thanh thông qua các bộ lọc Mel-frequency.
* **Áp dụng truyền thống Fourier rời rạc (DCT):** Sau đó, áp dụng biến đổi Fourier rời rạc (DCT) lên phổ năng lượng đã tính toán để chuyển đổi nó sang miền cepstral. Điều này tạo ra các hệ số cepstral dựa trên phổ năng lượng đã được biến đổi.
* **Tinh chỉnh và sử dụng các thông số đặc trưng:** Các hệ số cepstral thu được có thể được tinh chỉnh hoặc kết hợp với các thông số khác để tạo ra một biểu diễn đặc trưng cuối cùng của tín hiệu âm thanh.

HFCC chủ yếu được sử dụng trong các ứng dụng nhận dạng tiếng nói, *nơi mà sự chính xác và ổn định của đặc trưng đóng vai trò quan trọng* trong quá trình nhận dạng.

1. **Enhanced High-Frequency Cepstral Coefficients(HFCC-E):**

HFCC-E (Enhanced High-Frequency Cepstral Coefficients) là một biến thể của các hệ số cepstral cao tần số, được sử dụng rộng rãi trong xử lý tiếng nói và nhận dạng tiếng nói. HFCC-E cung cấp một biểu diễn đặc trưng cho tín hiệu âm thanh, tập trung vào các thông tin ở các tần số cao hơn, có thể cải thiện hiệu suất của các hệ thống nhận dạng tiếng nói, đặc biệt là trong môi trường nhiễu.

Cách tính HFCC-E thường tương tự như các bước để tính toán cepstral coefficients (MFCC hoặc HFCC) thông thường, nhưng có thể bao gồm các cải tiến hoặc thay đổi trong quá trình trích xuất đặc trưng. Cụ thể, HFCC-E *thường tập trung vào việc tăng cường thông tin ở các tần số cao hơn*, có thể bằng cách sử dụng bộ lọc hoặc biến đổi phức tạp hơn để chọn lọc thông tin.

HFCC-E có thể được sử dụng như là một phương pháp trích xuất đặc trưng thay thế hoặc bổ sung cho MFCC hoặc HFCC trong các ứng dụng nhận dạng tiếng nói và xử lý tín hiệu âm thanh, đặc biệt là trong môi trường có nhiều nhiễu hoặc ở các tần số cao hơn.

1. **Linear-frequency cepstral coefficients (LFCCs):**

Linear-frequency cepstral coefficients (LFCCs) là một biến thể của Mel-frequency cepstral coefficients (MFCCs), được sử dụng để trích xuất đặc trưng từ tín hiệu âm thanh. Trong LFCCs, thay vì sử dụng bộ lọc Mel để biến đổi tín hiệu âm thanh sang miền tần số, *ta sử dụng một bộ lọc tuyến tính với khoảng cách tần số tuyến tính*.

Quá trình tính toán LFCCs tương tự như tính toán MFCCs, bao gồm các bước chính sau:

* **Pre-emphasis:** Tương tự như MFCCs, áp dụng bộ lọc pre-emphasis để làm tôn lên các thành phần cao tần của tín hiệu âm thanh.
* **Framing:** Chia tín hiệu âm thanh thành các khung (frame) nhỏ có kích thước cố định.
* **Windowing:** Áp dụng cửa (window) như hàm Hamming cho mỗi frame để giảm thiểu hiện tượng "leakage" khi thực hiện biến đổi Fourier.
* **FFT (Fast Fourier Transform):** Thực hiện biến đổi Fourier nhanh (FFT) trên mỗi frame để chuyển đổi tín hiệu âm thanh từ miền thời gian sang miền tần số.
* **Linear Filterbank:** Tính toán năng lượng trong mỗi frame qua một bộ lọc tuyến tính, thay vì bộ lọc Mel như trong MFCCs.
* **Logarithm và DCT (Discrete Cosine Transform):** Tiếp theo, thực hiện các bước logarithm và DCT như trong MFCCs để thu được LFCCs.

LFCCs cũng được biểu diễn dưới dạng ma trận, với mỗi hàng tương ứng với một hệ số cepstral và mỗi cột tương ứng với một frame của tín hiệu âm thanh. LFCCs thường được sử dụng trong các ứng dụng như nhận dạng tiếng nói, nhận dạng người nói, và xử lý ngôn ngữ tự nhiên, tương tự như MFCCs.

1. **Relative Spectra Filtering (RASTA):**

Relative Spectra (RASTA) Filtering là một phương pháp trong xử lý tiếng nói để *cải thiện độ ổn định và ổn định của các đặc trưng âm thanh, đặc biệt là trong môi trường nhiễu.* Phương pháp này thường được sử dụng sau khi trích xuất các đặc trưng âm thanh như MFCC (Mel-Frequency Cepstral Coefficients) hoặc PLP (Perceptual Linear Prediction).

Cụ thể, RASTA Filtering thực hiện một loạt bước xử lý trên các đặc trưng âm thanh đã được trích xuất. Các bước này có thể bao gồm:

* **Trung bình động (Dynamic Mean Subtraction):** Áp dụng trung bình động để loại bỏ nhiễu từ tín hiệu âm thanh. Thay vì sử dụng trung bình của toàn bộ tín hiệu, trung bình động sử dụng một cửa sổ trượt để tính toán trung bình của các khung tín hiệu gần đó.
* **Dựa trên tần số (Frequency-Based Filtering):** Áp dụng bộ lọc trên tần số để cân bằng và giảm nhiễu tín hiệu. Các bộ lọc này thường được thiết kế để thực hiện một loạt phép toán trên các thành phần tần số của tín hiệu.
* **Mô hình ARMA (Auto-Regressive Moving-Average):** Sử dụng mô hình ARMA để ước lượng các hệ số và thực hiện filtering trên các đặc trưng âm thanh. Mô hình này giúp giảm bớt biến động và làm phẳng tín hiệu.

Mục tiêu của RASTA Filtering là *tăng cường độ ổn định và độ ổn định* của các đặc trưng âm thanh, giúp *cải thiện hiệu suất của các hệ thống nhận dạng tiếng nói trong môi trường nhiễu hoặc không ổn định*.

1. **RASTA-PLP:**

RASTA-PLP là một phương pháp *kết hợp giữa Relative Spectra (RASTA) Filtering và Perceptual Linear Prediction (PLP)* trong xử lý tiếng nói để trích xuất đặc trưng từ tín hiệu âm thanh. Phương pháp này kết hợp ưu điểm của cả hai kỹ thuật để cải thiện hiệu suất của hệ thống nhận dạng tiếng nói, đặc biệt là trong môi trường có nhiều nhiễu hoặc không ổn định.

Cụ thể, RASTA-PLP thực hiện các bước như sau:

* **Pre-emphasis:** Áp dụng bước pre-emphasis để tăng cường các tần số cao hơn của tín hiệu âm thanh.
* **Framing và Windowing:** Tín hiệu âm thanh được chia thành các khung nhỏ và áp dụng hàm cửa để giảm thiểu hiện tượng bị méo.
* **Linear Prediction (LP) Analysis:** Áp dụng phân tích Linear Prediction (LP) để ước lượng các hệ số dự đoán tuyến tính của tín hiệu.
* **RASTA Filtering:** Thực hiện RASTA Filtering trên các đặc trưng LP để cân bằng và giảm nhiễu tín hiệu, cải thiện độ ổn định của các đặc trưng.
* **Conversion to Mel Scale:** Chuyển đổi các tần số từ tần số Hertz sang tần số Mel sử dụng công thức đối với MFCC.
* **Discrete Cosine Transform (DCT):** Áp dụng biến đổi cosine rời rạc (DCT) cho các đặc trưng Mel sau RASTA Filtering để chuyển đổi từ miền tần số sang miền cepstral.
* **Cepstral Mean Normalization (CMN):** Thực hiện chuẩn hóa cepstral bằng cách trừ đi trung bình cepstral của tất cả các khung.

Kết quả của quá trình RASTA-PLP là một vectơ đặc trưng có chứa các hệ số cepstral đã được cải thiện bằng RASTA Filtering.

1. **Cepstral coefficients of Quasi-Cepestral Coefficients(CQCCs):**

Cepstral coefficients of Quasi-Cepestral Coefficients (CQCCs) là một phương pháp trích xuất đặc trưng từ tín hiệu âm thanh, tương tự như Mel-frequency cepstral coefficients (MFCCs) và Linear-frequency cepstral coefficients (LFCCs). Tuy nhiên, CQCCs được thiết kế để *cải thiện tính chịu biến đổi của các hệ số cepstral* khi tín hiệu âm thanh trải qua các biến đổi phổ biến như nhiễu, biến đổi kênh, và điều kiện môi trường khác nhau.

Quá trình tính toán CQCCs bao gồm các bước chính sau:

* **Gammatone Filterbank Analysis:** Tín hiệu âm thanh được đi qua một bộ lọc Gammatone để mô phỏng lựa chọn tần số của hệ thống cảm nhận âm thanh của con người.
* **Logarithmic Compression:** Năng lượng được tính toán từ bước trước được nén bằng cách lấy logarit.
* **Cepstral Analysis:** Cepstral analysis được thực hiện trên năng lượng đã được nén để thu được các hệ số cepstral.
* **Phase Information:** Thông tin về pha của tín hiệu cũng được giữ lại và được kết hợp với hệ số cepstral để tạo ra CQCCs.
* **Feature Representation:** Cuối cùng, các CQCCs được biểu diễn dưới dạng ma trận, với mỗi hàng tương ứng với một hệ số cepstral và mỗi cột tương ứng với một frame của tín hiệu âm thanh.

CQCCs thường được sử dụng trong các ứng dụng như nhận dạng tiếng nói, nhận dạng người nói, xử lý tín hiệu âm thanh với điều kiện *môi trường phức tạp*. Đặc biệt, chúng được đánh giá cao về tính ổn định và khả năng chịu biến đổi khi tín hiệu âm thanh trải qua các biến đổi phổ biến.

**[1]** [Microsoft Word - IJCAETS Kulkarni (arxiv.org)](https://arxiv.org/pdf/1305.1145)