BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

-----✵-----

TRẦN VĂN TÂM

XÁC ĐỊNH TẦN SỐ CƠ BẢN CỦA TÍN HIỆU TIẾNG NÓI DÙNG HÀM TỰ TƯƠNG QUAN

LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng – Năm 2019

BỘ GIÁO DỤC VÀ ĐÀO TẠO

ĐẠI HỌC ĐÀ NẴNG

-----✵-----

TRẦN VĂN TÂM

XÁC ĐỊNH TẦN SỐ CƠ BẢN CỦA TÍN HIỆU TIẾNG NÓI DÙNG HÀM TỰ TƯƠNG QUAN

Chuyên ngành: KHOA HỌC MÁY TÍNH

Mã số: 8480101

LUẬN VĂN THẠC SĨ KỸ THUẬT

Người hướng dẫn khoa học: **TS. NINH KHÁNH DUY**

Đà Nẵng – Năm 2019

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi.Các số liệu, kết quả nêu trong luận văn là trung thực. Mọi sự giúp đỡ cho việc thực hiện luận văn này đã được cảm ơn và các thông tin trích dẫn trong luận văn đã được chỉ rõ nguồn gốc rõ ràng và đượ phép công bố.

Người thực hiện luận văn

Trần Văn Tâm

LỜI CẢM ƠN

Sau thời gian học tập và rèn luyện, bằng sự biết ơn và kính trọng, tôi xin gửi lời cảm ơn chân thành đến Ban Giám hiệu, các phòng, khoa thuộc Trường đại học Đà nẵng và các Phó Giáo sư, Tiến sĩ đã nhiệt tình hướng dẫn, giảng dạy và tạo mọi điều kiện thuận lợi giúp đỡ tôi trong suốt quá trình học tập, nghiên cứu và hoàn thiện đề tài nghiên cứu khoa học này.

Đặc biệt, tôi xin bày tỏ lòng biết ơn sâu sắc tới TS Ninh Khánh Duy, người Thầy trực tiếp và cũng là người đã luôn tận tình hướng dẫn, chỉ bảo, giúp đỡ và động viên tôi trong suốt quá trình nghiên cứu và hoàn thành đề tài nghiên cứu này.

Xin chân thành cảm ơn gia đình, bạn bè cùng đồng nghiệp đã luôn khích lệ và giúp đỡ tôi trong quá trình học tập và nghiên cứu khoa học.

Người thực hiện luận văn

Trần Văn Tâm

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU

MỞ ĐẦU

1. Lý do chọn đề tài

Trong lịch sử phát triển của xã hội loài người, tiếng nói là một công cụ không thể thiếu. Tiếng nói giúp cho sự giao tiếp giữa con người và con người trở nên linh hoạt hơn, dễ hiểu nhau hơn. Tiếng nói chính là phương tiện để phân biệt con người với các loài động vật khác. Nhờ có tiếng nói, con người mới có xã hội, mới có sự phát triển đi lên qua nhiều hình thái xã hội

Trong lịch sử phát triển, chúng ta có nhiều hoạt động nghiên cứu liên quan đến tiếng nói nhằm để phục vụ lợi ích, nâng cao đời sống. Qua quá trình hoạt động nghiên cứu, chúng ta có đã có nhiều thành tựu trong lĩnh vực nghiên cứu tiếng nói. Và một trong những thành tựu quan trọng nhất của nghiên cứu tiếng nói đó là sự ra đời của điện thoại, khi mà âm thành không còn bị giới hạn bởi khoảng cách vật lý để chúng ta có thể truyền đạt thông tin cho nhau. Trải qua nhiều thế kỷ, các thành quả về nghiên cứu tiếng nói ngày càng trở nên quan trọng hơn với đời sống của chúng, và là một phần không thể thiếu trong cuộc sống hàng ngày.

Trong thời đại ngày nay, khi mà Công nghệ thông tin đang ngày càng góp phần quan trọng trong việc phục vụ lợi ích, nâng cao đời sống của chúng ta, việc áp dụng và mô phỏng tiếng nói cũng dần đóng vai trò quan trọng hơn. Nghiên cứu và mô phỏng tiếng nói cùng với trí tuệ nhân tạo đã và đang tạo thành xu thế và nghiên cứu chủ yếu trong giai đoạn này. Đặc biệt, khi công nghệ thông tin đang trở thành cốt lõi trong Cách mạng công nghiệp 4.0, việc nghiên cứu và mô phỏng tiếng nói dần trở nên quan trọng hơn, nhằm đưa máy móc gần với con người hơn trong việc giao tiếp giữa con người với con người, giữa máy móc với con người.

Một trong những tham số quan trọng trong lĩnh vực áp dụng và mô phỏng tiếng nói đó là tần số cơ bản F0. F0 là tần số cơ bản của tín hiệu tiếng nói (đơn vị Herz).Về âm học tần số cơ bản chính là F0 tốc độ rung của dây thanh (vocal cord) của bộ máy phát âm của con người [1]. Về cảm nhận âm thanh, F0 tương quan với cao độ (độ trầm bổng) của lời nói (F0 càng cao thì giọng nói càng bổng).

F0 rất quan trọng để nắm bắt và xử lý tiếng nói cho các nghiên cứu sâu hơn. Nghiên cứu và hiểu rõ được tần số cơ bản F0 có thể là cơ sở cho các nghiên cứu ứng dụng khác.

Trong lĩnh vực phân tích tiếng nói, tính F0 được ứng dụng trong việc đo cao độ trung bình của một người, biểu diễn ngữ điệu của lời nói dựa trên tín hiệu thu được. Trong tổng hợp tiếng nói, việc tính F0 là cơ sở để máy tính tái tạo tiếng nói có đặc tính ngữ điệu giống với tiếng nói tự nhiên. Trong nhận dạng tiếng nói, việc tính tần số cơ bản F0 giúp tăng tỷ lệ nhận dạng đúng nếu kết hợp thêm đặc trưng ngữ điệu. Ngoài ra, bài toán tính F0 có nhiều ứng dụng khác như: máy móc nhận diện giọng nói của con người để thực hiện lệnh, máy móc nhận diện được thái độ trong tiếng nói để xác định tâm trạng của con người,...

Để xác định được tần số cơ bản F0 của tiếng nói thì được chia thành hai nhóm: các thuật toán trên miền thời gian (time domain) và các thuật toán trên miền tần số (frequency domain) [4]. Trong phạm vi của luận văn, tôi nghiên cứu thuật toán trên miền thời gian, sử dụng hàm tự tương quan (autocorelation) [2][4], đồng thời kết hợp với thuật toán lọc trung vị để làm trơn kết quả F0 thu được. Qua đó đánh giá thuật toán trên với cách tính thủ công để tìm F0.

2. Mục đích và ý nghĩa đề tài

a. Mục đích

Mục đích nghiên cứu đề tài:

- Nghiên cứu và cài đặt thuật toán tính tần số cơ bản F0 của tín hiệu tiếng nói trên miền thời gian dùng hàm tự tương quan.
- Phân tích ưu nhược điểm của thuật toán tự tương quan tính F0 trên miền thời gian.
- Khảo sát tác dụng của lọc trung vị nhằm làm trơn kết quả tính F0 tự động.
- So sánh và đánh giá giữa hai phương pháp tính F0: dùng hàm tự tương quan, và thủ công.

b. Ý nghĩa khoa học và thực tiễn của đề tài

- Đóng góp phương pháp tính tần số cơ bản F0 trong lĩnh vực xử lý tín hiệu tiếng nói.
- Đưa ra kết quả khi áp dụng trong thực tế đối với hai hàm xác định tần số cơ bản F0, là cơ sở cho các nghiên cứu, đánh giá để tính tần số cơ bản F0 sau này.

3. Mục tiêu và nhiệm vụ

a. Mục tiêu

Mục tiêu chính của đề tài là nghiên cứu phương pháp tính tần số cơ bản F0 dựa trên hàm tự tương quan, lọc trung vị, và phân tích ưu nhược điểm của các thuật toán.

b. Nhiệm vụ

Để đạt được mục tiêu, nhiệm vụ đặt ra của đề tài là:

- Nghiên cứu lý thuyết liên quan đến tần số cơ bản F0.
- Nghiên cứu lý thuyết hàm tự tương quan.
- Thực hiện phân tích, đánh giá kết quả tính F0, kết hợp với làm trơn kết quả qua thuật toán lọc trung vị.

4. Đối tượng và phạm vi nghiên cứu

a. Đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài là tín hiệu tiếng nói và các thuật toán xử lý tín hiệu tiếng nói.

b. Phạm vi nghiên cứu

Phạm vi nghiên cứu của đề tài là các thuật toán tính F0 của tín hiệu tiếng nói trên miền thời gian.

5. Phương pháp nghiên cứu

a. Phương pháp lý thuyết

- Thu thập và nghiên cứu các tài liệu liên quan đến đề tài.

b. Phương pháp thực nghiệm

Nghiên cứu và khai thác các công cụ, phần mềm hỗ trợ.

- So sánh, thử nghiệm, đánh giá kết quả tính F0 dựa trên hai phương pháp tính là tự tương quan kết hợp với lọc trung vị làm trơn kết quả.
- So sánh, đánh giá kết quả của thuật toán dùng tự tương quan tính tần số cơ bản F0 với cách đo thủ công.

6. Kết luận

a. Kết quả của đề tài

- Nghiên cứu và tính được tần số cơ bản F0 dựa trên thuật toán dùng tự tương quan.

- Đánh giá sai số của thuật toán dùng hàm tự tương quan tính F0 tự động dựa trên kết quả đo tần số cơ bản F0 thủ công.

b. Hướng phát triển của đề tài

- Nghiên cứu giải pháp để cải thiện độ chính xác của các thuật toán tính tần số cơ bản F0 trên miền thời gian.
- Đề xuất, cải tiến để thực hiện tính F0 theo thời gian thực.

7. Bố cục của luận văn

Dự kiến luận văn được trình bày bao gồm các phần chính như sau:

MỞ ĐẦU

Nêu bối cảnh nghiên cứu, lý do chọn đề tài và mục tiêu nghiên cứu.

CHƯƠNG I: TỔNG QUAN VỀ XỬ LÝ TÍN HIỆU TIẾNG NÓI

Trong chương này trình bày các khái niệm cơ bản của tiếng nói, quá trình hình thành tiếng nói và các đặc tính cơ bản của tín hiệu tiếng nói.

CHƯƠNG II: THUẬT TOÁN TÌM F0 CỦA TÍN HIỆU TIẾNG NÓI

Trong chương này trình bày lý thuyết về hàm tự tương quan. Ngoài ra, do đề tài có sử dụng thuật toán lọc trung vị để làm trơn kết quả nên thuật toán này cũng được nêu trong chương này.

CHƯƠNG III: TRIỂN KHAI VÀ ĐÁNH GIÁ CÁC THUẬT TOÁN

Để áp dụng được các thuật toán trên Matlab, trong chương này trình bày công cụ Matlab và các hàm liên quan đến xử lý tín hiệu tiếng nói [5][6].

Trong chương này thực hiện áp dụng hai hàm tự tương quan tự triển khai, hàm tự tương quan của công cụ Matlab để tính F0. Đồng thời, kết hợp với thuật toán lọc trung vị để làm trơn kết quả.

Trong chương này cũng đưa ra so sánh giữa các phương pháp, so sánh với các kết quả tính F0 thủ công và đánh giá từ đó rút ra được ưu nhược điểm của hàm tự tương quan trong việc tính tần số cơ bản F0.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

PHỤ LỤC

TÀI LIỆU THAM KHẢO

CHƯƠNG 1: TỔNG QUAN VỀ XỬ LÝ TÍN HIỆU TIẾNG NÓI

1.1. Mở đầu

Trong giao tiếp của con người, tiếng nói như là một phương tiện giao tiếp cơ bản và nhanh nhất để biểu đạt ý của người muốn truyền đạt. Xã hội chúng ta phát triển cũng là nhờ có tiếng nói để có thể truyền đạt ý kiến, mong muốn giữa người với người. Để hỗ trợ cho việc giao tiếp bằng tiếng nói, con người có thể dùng các cử chỉ, điệu bộ của chân tay làm cho các ý muốn truyền đạt nhanh hơn đến người muốn truyền đạt. Vì là giao tiếp trực tiếp nên tiếng nói là phương thức truyền đạt nhanh nhất giữa những người muốn giao tiếp với nhau. Sở dĩ như vậy, ngoài tiếng nói còn có chữ viết để con người có thể giao tiếp với nhau. Tuy nhiên, chữ viết là phương thức truyền đạt gián tiếp nên sẽ chậm hơn phương thức truyền đạt là tiếng nói. Với sự phát triển của công nghệ, để có sự giao tiếp trở nên linh hoạt hơn, tiếng nói như là một công cụ hỗ trợ mạnh mẽ để thúc đẩy việc biểu diễn tiếng nói trong khoa học máy tính. Tiếng nói được sử dụng như là một dữ liệu được lưu trữ trong máy tính, qua đó có thể truyền đạt thông qua mạng truyền thông để phục vụ nhiều mục đích khác nhau để phục vụ lợi ích trong đời sống của con người. Trong các hệ thống xử lý tiếng nói, cần chú ý đến hai điểm: sự nguyên vẹn của nội dung thông điệp trong tin hiệu tiếng nói; biểu diễn tín hiệu tiếng nói phải tiện lợi cho việc truyền tải, lưu trữ hoặc trong một dạng linh động để có thể chuyển đổi thành tín hiệu liếng nói mà không giảm nội dung của thông điệp [4].

1.2. Khái niệm về tín hiệu tiếng nói

Con người có năm giác quan để cảm nhận và nhận thức thế giới xung quanh. Trong quá trình phát triển của xã hội loài người, con người dùng năm giác quan này để nhận thức, thu thập kiến thức và tác động trở lại tự nhiên qua đó nâng cao đời sống của con người. Một trong những giác quan quan trọng trong sự phát triển của xã hội con người đó là thính giác. Nhờ có thính giác mà con người có thể nghe được âm thanh, con người có thể giao tiếp được với nhau qua âm thanh.

Về bản chất, âm thanh từ lời nói, âm thanh trong thế giới tự nhiên đều là những sóng âm lan truyền trong môi trường. Khi chúng ta nói dây thanh trong hầu bị chấn động, tạo nên những sóng âm, sóng truyền trong không khí đến màng nhĩ – một màng mỏng rất

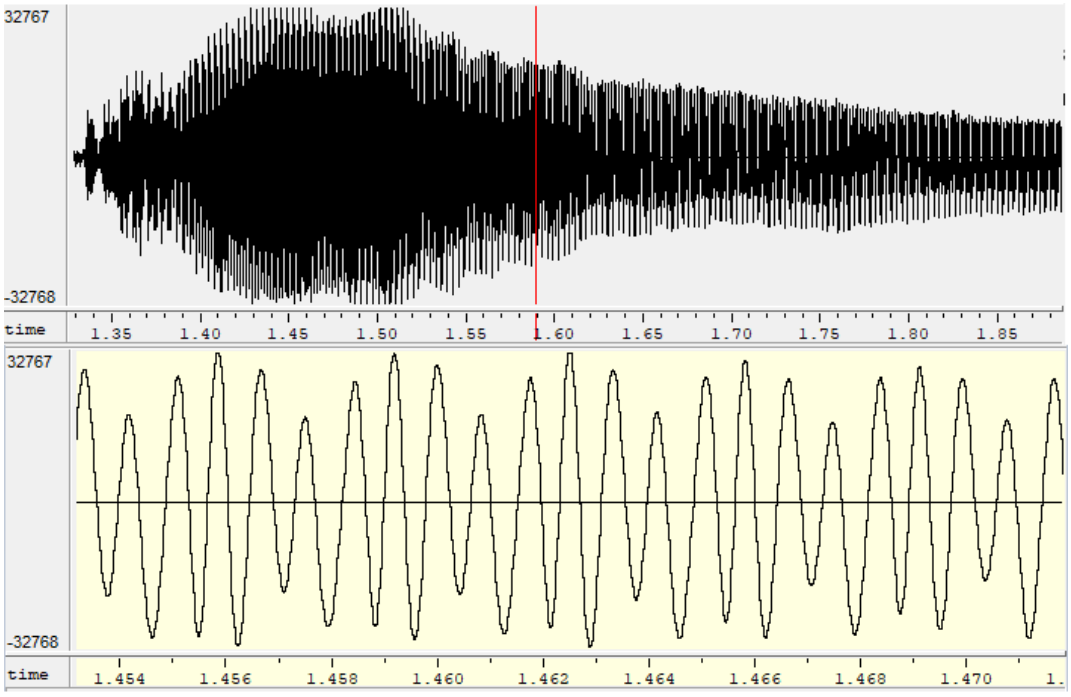
nhạy cảm của tai ta – làm cho màng nhĩ cũng dao động, các dây thần kinh của màng nhĩ sẽ nhận được cảm giác âm khi tần số dao động của sóng đạt đến một độ lớn nhất định.

Tai con người chỉ cảm thụ được những dao động có tần số từ khoảng 16 Hz đến khoảng 20000 Hz. Những dao động trong miền tần số này gọi là dao động âm hay âm thanh, và các sóng tương ứng gọi là sóng âm. Những sóng có tần số nhỏ hơn 16 Hz gọi là sóng hạ âm, những sóng có tần số lớn hơn 20000 Hz gọi là sóng siêu âm, con người không cảm nhận được (ví dụ loài dơi có thể nghe được tiếng siêu âm) [1].

Tất cả các sóng âm đều được lan truyền trong môi tường, từ môi trường không khí, môi trường rắn, môi trường lỏng.

Trong xử lý tín hiệu tiếng nói, tín hiệu tiếng nói có hai cách để biểu diễn: biểu diễn tín hiệu trên miền thời gian và biểu diễn tín hiệu trên miền tần số.

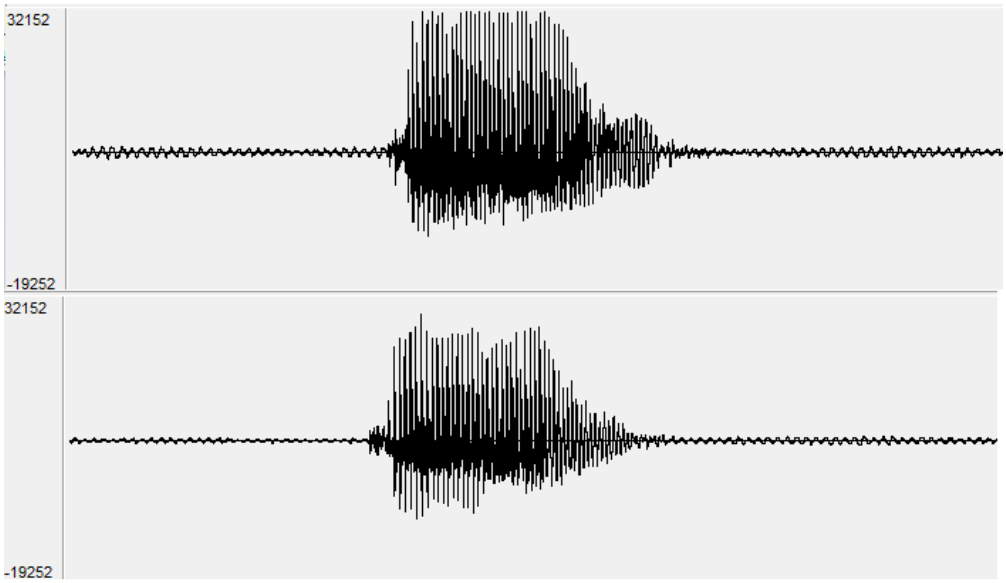
1.2.1. Biểu diễn trên miền thời gian



Hình 1.1 – Dạng sóng theo thời gian

Âm thanh dưới dạng sóng được lưu trữ theo định dạng thông dụng trong máy tính là file .wav với các tần số lấy mẫu thường gặp là: 8000 Hz, 10000 Hz, 11025 Hz, 16000 Hz, 22050 Hz, 32000 Hz, 44100 Hz,…; độ phân giải hay còn gọi là số bít/mẫu là 8 hoặc 16 bít và số kênh là 1 (Mono) hoặc 2 (Stereo).

Tuỳ theo thiết bị, thời điểm, người phát âm thì dữ liệu âm thanh được số hoá, biểu diễn lại trong máy tính sẽ khác nhau.



Hình 1.2 – Tín hiệu của cùng một âm do một người nói thu ở hai thời điểm khác nhau

1.2.2. Biển diễn trên miền tần số

Một trong những đại lượng đặc trưng để biểu diễn tín hiệu tiếng nói trên miền tần số đó là phổ.

Phổ trong tín hiệu tiếng nói là biểu diễn của sự phụ thuộc của biên độ vào thời gian và tần số, là hình ảnh biểu diễn của tín hiệu tiếng nói theo trục của tần số.

1.2.2.1. Biến đổi Fourier

Biến đổi Fourier trong xử lý tín hiệu tiếng nói là phép biến đổi tín hiệu tiếng nói theo miền thời gian sang miền tần số.

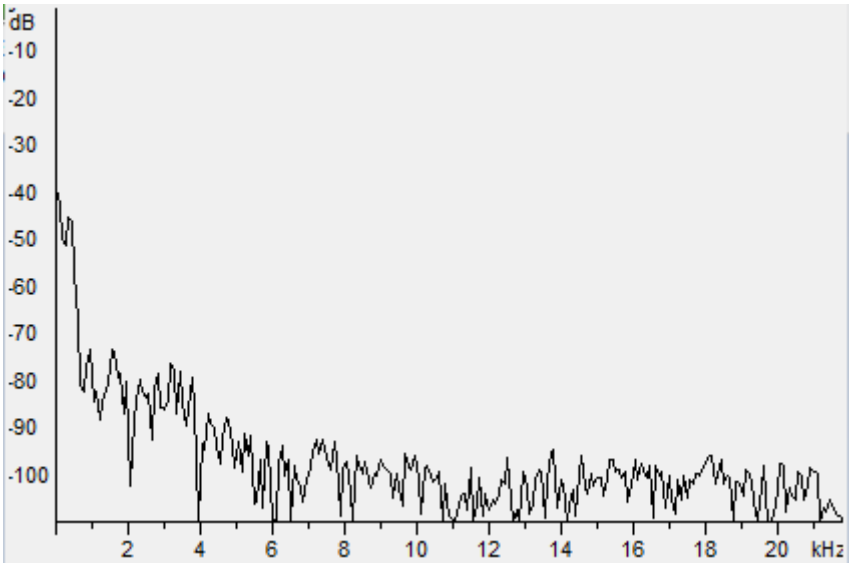
Biến đổi Fourier có nhiều dạng:

- Biến đổi Fourier liên tục là một toán tử tuyến tính chuyển một hàm tích phân này sang một hàm tích phân khác. Trong xử lý tín hiệu, biến đổi Fourier liên tục được áp dụng trên phổ và theo các thành phần trong phổ.

- Biến đổi Fourier rời rạc là phép biến đổi cho các tín hiệu thời gian rời rạc. Biến đổi này thương được áp dụng trong việc phân tích phổ, lọc tín hiệu.

1.2.2.2. Phổ hai chiều

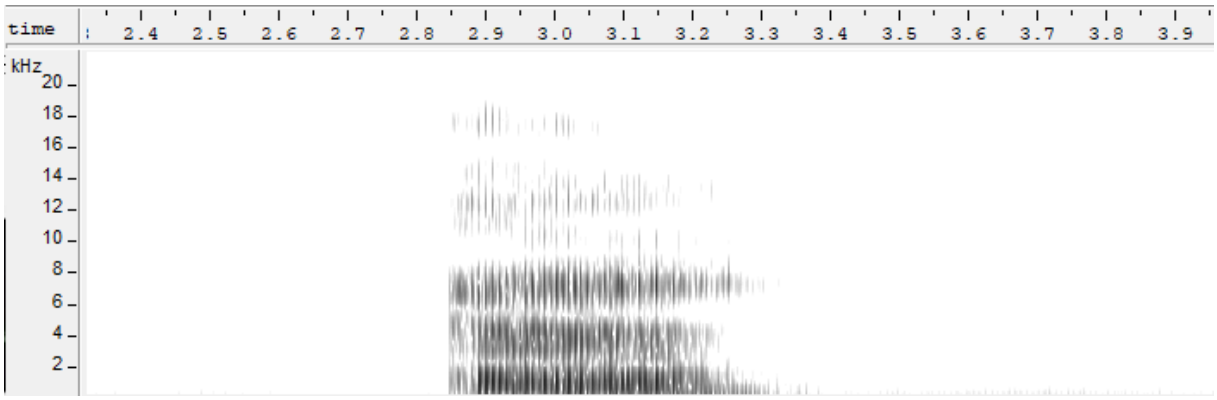
Phổ hai chiều là phổ trong đó chứa thông tin tin hiệu tiếng nói và được biểu diễn trên hai đại lượng là tần số và biên độ phổ.



Hình 1.3 – Phổ hai chiều

1.2.2.3. Phổ ba chiều

Phổ ba chiều là phổ trong đó tín hiệu tiếng nói được biểu diễn trên ba đại lượng: thời gian, tần số, và biên độ phổ.



Hình 1.4 – Phổ ba chiều

Nếu màu của tín hiệu càng đậm thì biên độ phổ (hay năng lượng của tín hiệu) càng cao.

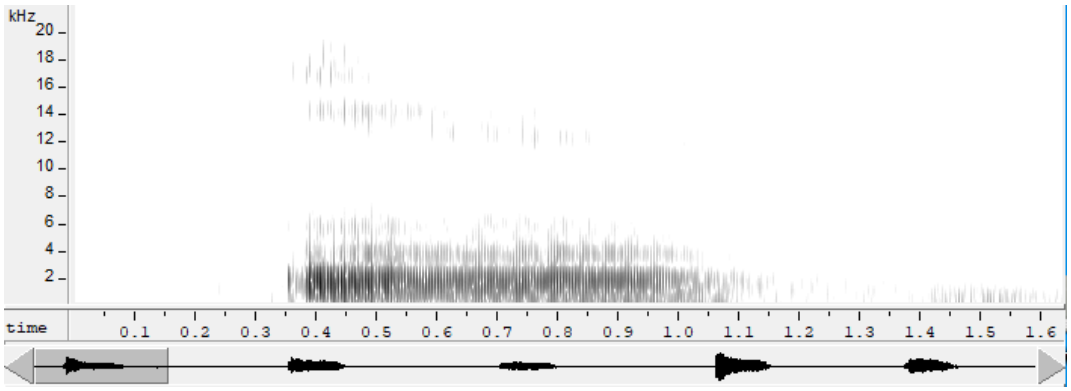
1.3. Các đặc tính cơ bản của tín hiệu tiếng nói

Tiếng nói được tạo ra từ độ rung của dây thanh âm trong thanh quản thông qua khí quản và hoạt động của tuyến âm. Như vậy, tiếng nói chính là âm thanh. Tiếng nói có chu kỳ dao động, có tần số âm thanh.

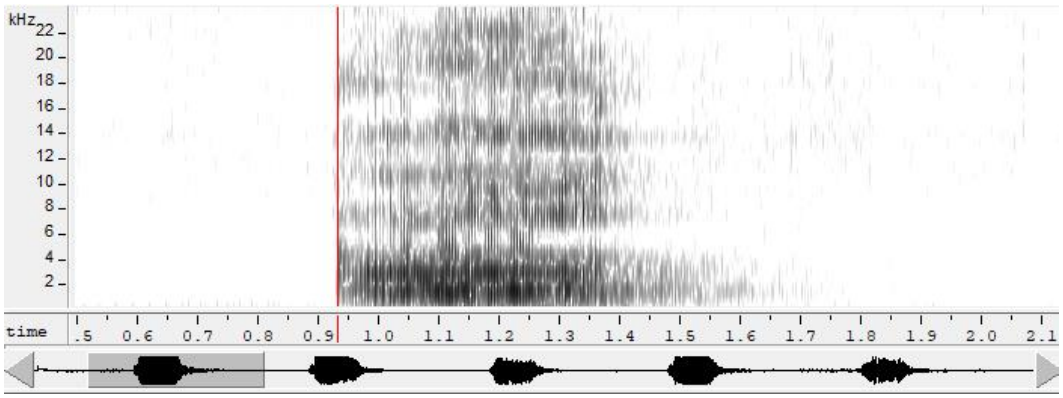
1.3.1. Âm sắc

Âm sắc là một trong bốn đặc tính cơ bản của âm thanh cũng như tín hiệu tiếng nói. Âm sắc giúp ta phân biệt được tiếng nói của từng âm và của mỗi người được cảm nhận khác nhau như thế nào. Âm sắc liên quan mật thiết đến phổ của tín hiệu.

Hình dưới đây minh hoạ âm sắc (dưới dạng phổ 3 chiều) ứng với nữ giới và nam giới khi phát cùng một âm.



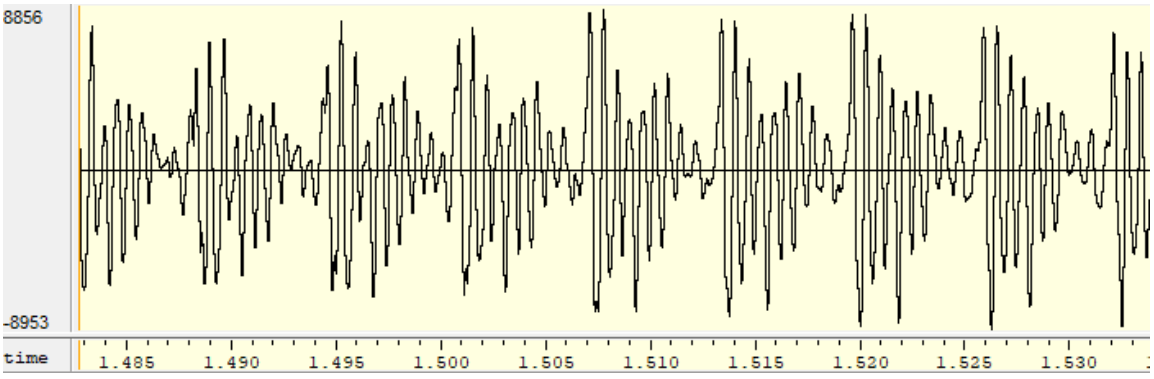
Hình 1.5 – Âm sắc của một người nữ khi phát nguyên âm /a/



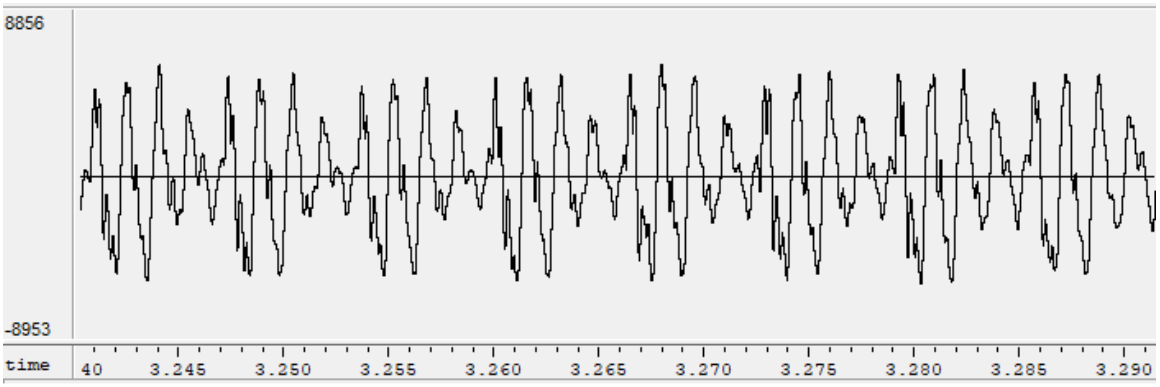
Hình 1.6 -Âm sắc của một người nam khi phát nguyên âm /a/

1.3.2. Cường độ

Cường độ là độ to hay nhỏ của âm thanh nói ra. Cường độ càng lớn thì âm thanh truyền càng xa trong môi trường truyền. Cường độ âm là số năng lượng mà sóng âm truyền đi trong một thời gian nhất định trên đơn vị diện tích cố định và vuông góc với phương truyền âm. Trong tiếng nói, cường độ của nguyên âm thường lớn cường độ của phụ âm. Trên đồ thị biểu diễn sóng tín hiệu (waveform), cường độ âm thanh tỉ lệ thuận với giá trị tuyệt đối của biên độ tín hiệu.



Hình 1.7 – Đồ thị biểu diễn sóng tín hiệu của nguyên âm /a/ của một người nói

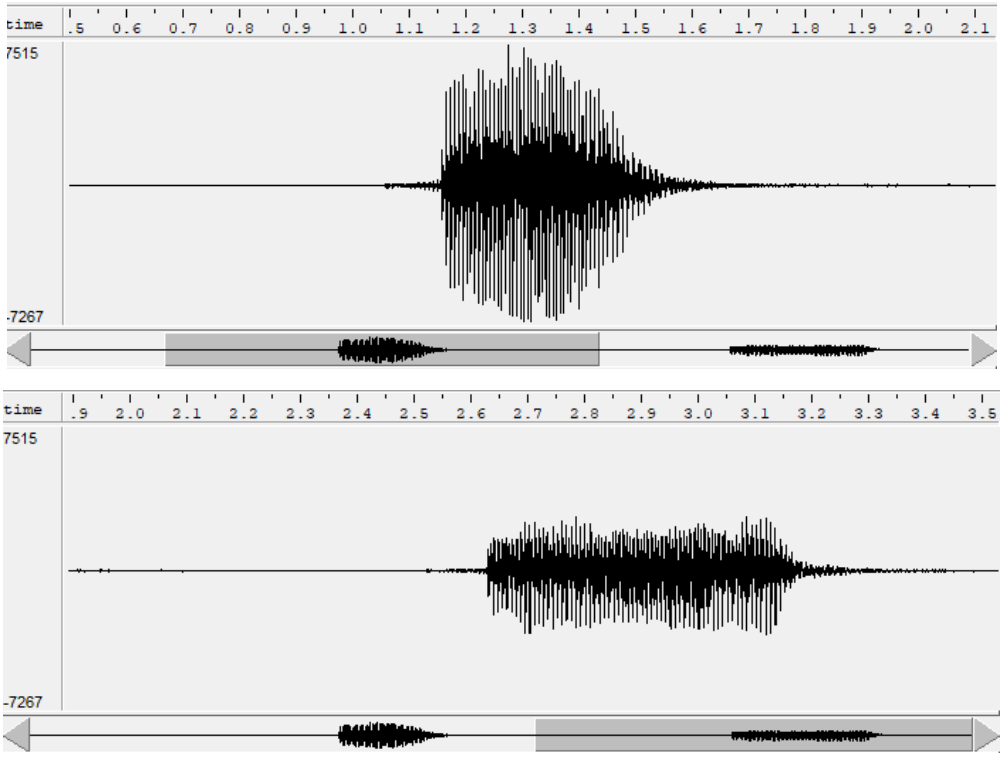


Hình 1.8 – Đồ thị biểu diễn sóng tín hiệu của phụ âm /h/ của một người nói

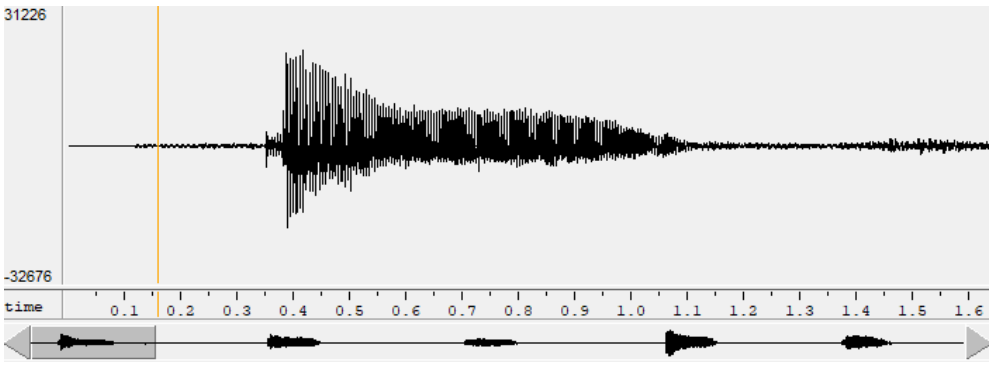
1.3.3. Trường độ

Trường độ hay còn được biết là độ dài của âm phát ra phụ thuộc vào sự chấn động lâu hay nhanh của phần tử môi trường truyền đi.

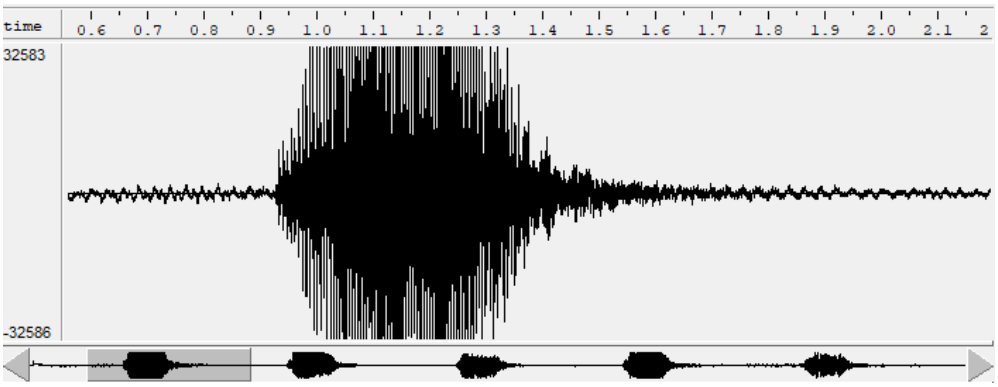
Trường độ của mỗi người khác nhau và mỗi thời điểm cũng khác nhau.



Hình 1.9 – Nguyên âm /a/ được thu ở hai thời điểm khác nhau của cùng một người nói



Hình 1.10 – Âm /a/ của một người nữ



Hình 1.11 – Âm /a/ của một người nam

1.3.5. Âm hữu thanh

Âm hữu thanh (voiced speech) là âm phát ra có thanh, ví dụ như các nguyên âm /a/, /e/, /i/, /o/, /u/ hoặc các phụ âm như /m/, /n/, /l/. Thực ra âm hữu thanh được tạo ra là do việc không khí qua thanh môn (thanh môn tạo ra sự khép mở của dây thanh dưới sự điều khiển của hai sụn chóp) với một độ căng của dây thanh sao cho chúng tạo nên dao động.

Trong xử lý tín hiệu tiếng nói, âm hữu thanh gồm các khung tín hiệu tuần hoàn nên có thể tính được tần số cơ bản F0.

1.3.6. Âm vô thanh

Âm vô thanh (voiced speech) là âm khi tạo ra tiếng thì dây thanh không rung hoặc rung đôi chút tạo ra giọng như giọng thở, ví dụ như /t/, /p/ hay /k/.

Trong xử lý tín hiệu tiếng nói, âm vô thanh không có ích khi tính tần số cơ bản. Vì âm vô thanh không có khung tín hiệu tuần hoàn. Tần số cơ bản ở âm vô thanh là không xác định.

1.4. Xử lý ngắn hạn (short-time processing)

Tín hiệu tiếng nói có một tính chất quan trọng là các đặc tính của nó thay đổi tương đối chậm theo thời gian. Thông thường, các đặc tính của tín hiệu ổn định trong khoảng

thời gian từ 10 ms đến 30 ms. Do đó, người ta thường chia tín hiệu cần xử lý thành các khung tín hiệu liên tiếp nhau, mỗi khung có độ dài từ 10 ms đến 30 ms. Sau đó, ta tiến hành xử lý trên mỗi khung tín hiệu này. Các khung tín hiệu này được gọi là các khung phân tích, các khung này có thể trùng nhau (overlap) một phần để đảm bảo các đặc tính của tín hiệu biến đổi trơn tru giữa 2 khung liên tiếp. Việc chia khung này sẽ được lặp lại từ đầu đến cuối trên tín hiệu cần xử lý. Kết quả của việc xử lý trên mỗi khung có thể chỉ gồm một giá trị số (ví dụ như giá trị năng lượng hoặc giá trị F0), có thể gồm nhiều giá trị số (ví dụ như các hệ số phổ).

Hình 1.12 – Chia tín hiệu thành các khung cửa sổ

Việc chia tín hiệu tiếng nói thành các khung tín hiệu giúp ta xác định và xử lý được các tín hiệu tiếng nói có đặc tính hầu như không thay đổi, độc lập.

Hầu hết các kỹ thuật xử lý ngắn hạn được biểu diễn dưới dạng:

$$Q_n = \sum_{m=-\infty}^{\infty} T[x(m)]w(n-m) \qquad (1.1)$$

Tín hiệu tiếng nói được biến đổi bởi hàm *T[ ]*, tuyến tính hoặc phi tuyến tính, và có thể phụ thuộc vào một vài điều chỉnh thông số hoặc tập các thông số. Kết quả là các cửa sổ có trình tự và vị trí, thời gian tương ứng với mẫu chỉ số n. Và kết quả là tổng giá trị các số khác không. Thông thường, các cửa số tuần tự này có thời gian giới hạn. Giá trị Q_n là tuần tự các trọng số trung bình của trình tự *T[x(m)]*

Năng lượng ngắn hạn của tín hiệu tiếng nói là ví dụ đơn giản minh hoạ cho ý tưởng ở trên.

$$E = \sum_{m=-\infty}^{\infty} x^2(m) \qquad (1.2)$$

Tuy nhiên, đại lượng trên có ít ý nghĩa với các thông tin về các thuộc tính phụ thuộc thời gian trong tín hiệu tiếng nói. Nên đại lượng trên được đơn giản lại

$$E_n = \sum_{m=n-N+1}^{n} x^2(m) \quad (1.3)$$

Năng lượng thời gian ngắn hạn tại mẫu n là tổng bình phương của N mẫu từ n – N + 1 đến n.

Với

$$w(n) = 1 \text{ với } 0 \le n \le N\text{-}1$$
$$= 0 \text{ trong trường hợp khác}$$

Biên độ của tín hiệu tiếng nói thay đổi đáng kể theo thời gian. Hầu hết trong các trường hợp, âm vô thanh có biên độ thấp hơn đối với các âm hữu thanh. Năng lượng ngắn hạn của tín hiệu tiếng nói phản ánh những biên độ dao động. Ta có thể định nghĩa lại năng lượng ngắn hạn như sau:

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \quad (1.4)$$

Biểu thức trên được viết lại

$$E_n = \sum_{m=-\infty}^{\infty} x(m)^2 \cdot h(n-m) \quad (1.5)$$

với

$$h(n) = w^2(n) \quad (1.6)$$

Tín hiệu $x^2(n)$ được lọc bởi bộ lọc tuyến tính với đáp ứng xung $h(n)$

Có trường hợp với N tăng lên, các dao động biên độ không thay đổi, năng lượng ngắn hạn cũng không thay đổi, hoặc ít thay đổi. Vì vậy, đối với cửa sổ với khung thời gian ngắn quá thì cũng không cung cấp đủ thông tin về thay đổi biên độ của tín hiệu tiếng nói.

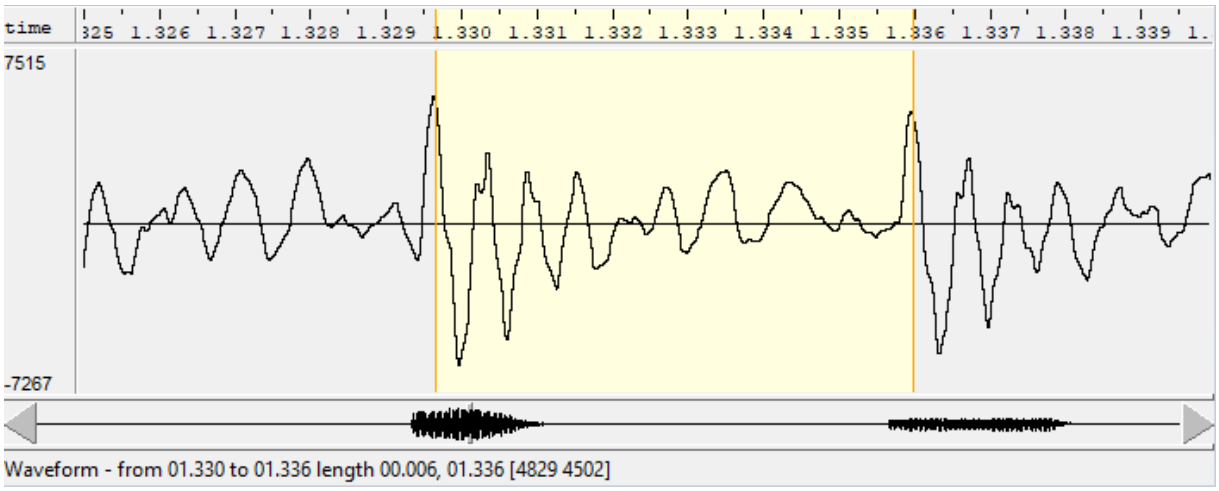
Nếu N quá nhỏ thì năng lượng quá hạn E_n sẽ dao động nhanh tuỳ thuộc vào chi tiết chính xác của dạng sóng. Nếu N quá lớn, E_n sẽ thay đổi rất chậm, vì vậy sẽ không phản ánh được sự thay đổi của thuộc tính tín hiệu tiếng nói.

Trong thực tế, thời lượng của chu kỳ cao độ thay đổi từ 20 mẫu (tại tốc độ lấy mẫu 10 kHz) với cao độ nữ và với 250 mẫu đối với cao độ nam nên không có giá trị đơn nào của N đáp ứng được. Vì vậy, N sẽ được chọn theo thứ tự từ 100 đến 200 mẫu cho tốc độ lấy mẫu 10 kHz (từ 10 đến 20 ms).

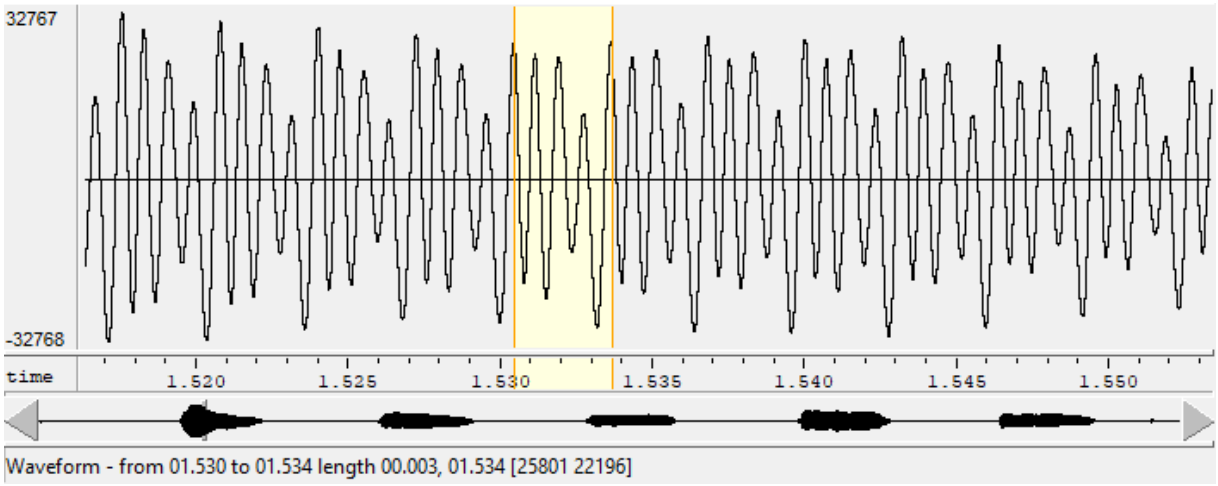
1.5. Tần số cơ bản (F0)

1.5.1. F0 là gì

Tần số cơ bản là tốc độ rung của dây thanh trong quá trình phát âm, gọi là F0. Người nói có thể điều khiển mức độ căng của hai dây thanh để khoảng giữa hai dây thanh đó đóng lại hoàn toàn, tạo thành khe hẹp hay mở rộng ra. Khoảng không ở giữa này được gọi là thanh môn. Khi thanh môn hẹp, không khí đi qua nó sẽ tạo ra một âm thanh điều hòa. Thuật ngữ "cao độ" (pitch) dùng để chỉ tần số cơ bản mà người nghe có thể cảm nhận được. Bằng cách thay đổi độ căng của dây thanh, người nói có thể điều chỉnh tần số cơ bản. Thông thường, F0 của giọng nam nằm trong khoảng từ 70 Hz đến 250 Hz, trong khi đó giọng nữ có F0 từ 150 Hz đến 400 Hz [4].



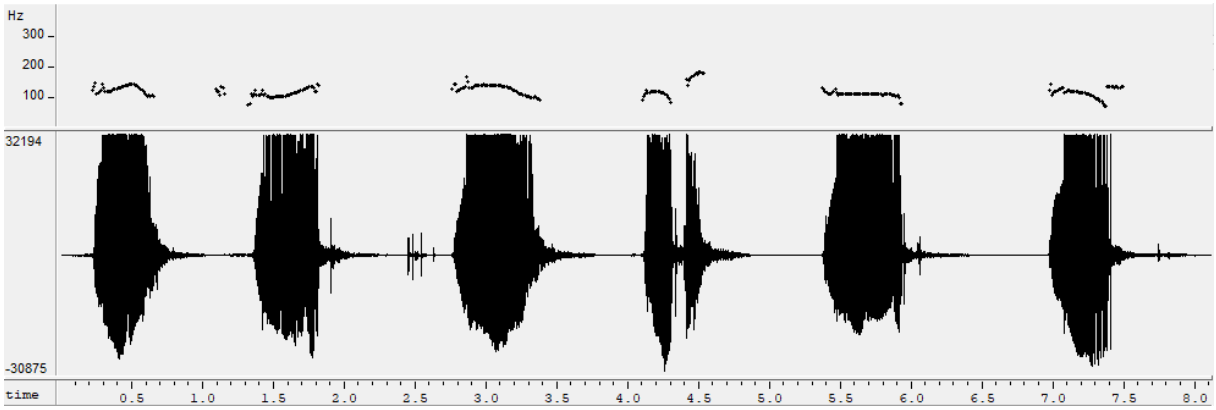
Hình 1.13 – Tần số cơ bản đo ở nguyên âm /a/ của một người nam là 166.6 Hz ứng với chu kỳ cơ bản là 0.006 giây



Hình 1.14 – Tần số cơ bản đo ở nguyên âm /a/ của một người nữ là 333.3 Hz ứng với chu kỳ cơ bản là 0.003 giây

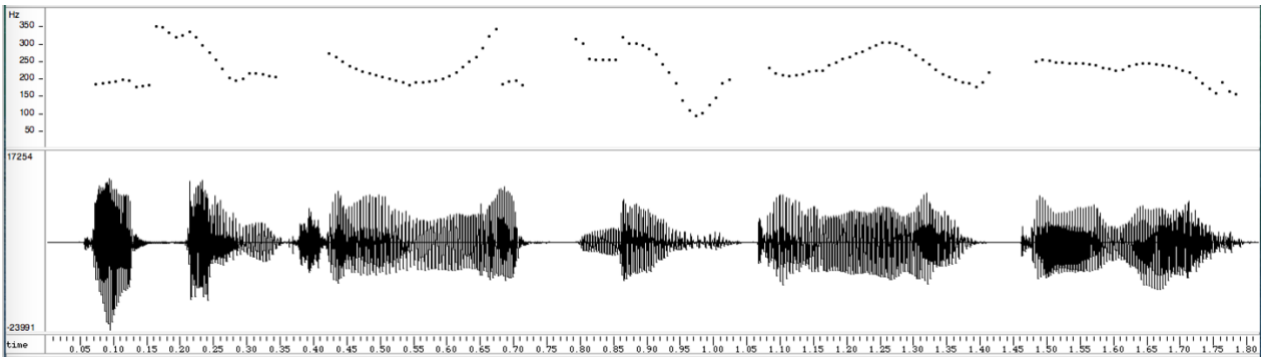
1.5.2. Tầm quan trọng của F0 trong xử lý tiếng nói

Trong xử lý tiếng nói, F0 đặc trưng cho ngữ điệu của lời nói (đặc trưng chung cho mọi ngôn ngữ) và thanh điệu của âm tiết (đặc trưng riêng cho tiếng Việt). Đây là hai tham số quan trọng của tiếng nói. Việc xác định F0 có các ứng dụng trong nhận dạng tiếng nói và tổng hợp tiếng nói. Nhận dạng chính xác thanh điệu của mỗi âm tiết giúp cải thiện hiệu năng của hệ thống nhận dạng tiếng nói [3]. Trong tổng hợp tiếng nói, việc mô hình hoá chính xác đường F0 của mỗi thanh điệu giúp máy tính sinh ra tiếng nói tự nhiên hơn [9].



Hình 1.15 – Đường F0 của các thanh điệu tiếng Việt

Hình 1.15 cho thấy đường F0 được xác định qua các âm được thu thành file .wav trong điều kiện phòng. Đoạn tín hiệu trên là phát âm của một người nam phát âm chữ "ba", "bá", "bà", "bã", "bạ", "bả". Chữ "ba" cho thấy F0 là dãy ít thay đổi giá trị (thanh bằng), chữ "bá" có giá trị F0 tăng dần, chữ "bà" cho thấy F0 có giá trị giảm dần, chữ "bã" cho thấy giá trị F0 có sự gián đoạn, chữ "bạ" cho thấy F0 có giá trị đồng đều rồi giảm đột ngột, chữ "bả" có F0 tương tự như chữ "bã" nhưng dãy giá trị F0 đoạn thứ hai ít thay đổi giá trị. Như vậy, qua giá trị F0 tính được, có thể suy diễn được thanh điệu của âm tiết phát ra trong một đoạn tín hiệu tiếng nói.



Hình 1.16 – Đường F0 (trên) và tín hiệu (dưới) của câu nói "Các bạn trẻ nhất định có nhiều cơ hội" của một giọng nữ

Hình 1.16 minh hoạ một ví dụ về tín hiệu của một đoạn câu nói được thu âm lại và đường F0 đo được. Qua hình trên, các giá trị F0 cho thấy sự thay đổi của ngữ điệu trong câu nói, ngữ điệu có đoạn đi lên và có đoạn đi xuống trong quá trình nói.

1.5.3. Các lý do khiến việc tìm F0 khó khăn

Có nhiều nguyên nhân khiến cho việc xác định F0 của tín hiệu tiếng nói khó khăn [4]. Ở đây tôi tóm lại có 3 nguyên nhân chính sau.

Một là, tín hiệu tiếng nói về bản chất là tín hiệu ngẫu nhiên, không theo quy luật nhất định, dẫn đến việc tìm quy luật về tính tuần hoàn của tín hiệu tiếng nói không dễ dàng. Tính ngẫu nhiên thể hiện ở chỗ tín hiệu tiếng nói thu được của cùng một âm thay đổi theo rất nhiều yếu tố bao gồm: điều kiện thu âm (thiết bị thu, khoảng cách từ thiết bị thu đến miệng người nói, môi trường thu âm), người nói, thời điểm thu âm, thể trạng (điều kiện tâm lý và sức khoẻ) của người nói tại thời điểm thu âm,… Một số ví dụ về tính ngẫu nhiên của tín hiệu tiếng nói đã được trình bày trong các phần trước của chương này.

Hai là, trong môi trường thu âm thực tế, không chỉ có tiếng nói mà còn các nguồn âm khác được phát ra. Do đó, tín hiệu được thu lại ngoài tiếng nói còn có những âm thanh khác được thu vào. Những âm thanh này gọi là các tạp âm (hay nhiễu). Nhiễu lẫn vào tín hiệu tiếng nói sẽ làm cho thuật toán xử lý bị sai lệch. Các ví dụ điển hình là: nhiễu có biên độ lớn làm méo hình dạng của tín hiệu tiếng nói gốc, hoặc nhiễu có thể vô tình có dạng sóng tuần hoàn dẫn đến thuật toán tính F0 tưởng nhầm là âm hữu thanh để đi tính F0 một cách không cần thiết.

Ba là, trong các cơ quan phát âm đóng góp vào việc tạo nên tiếng nói, ngoài dây thanh (liên quan đến tính tuần hoàn hay F0 của tín hiệu) còn có khoang miệng và khoang mũi (liên quan đến hình dạng chung hay âm sắc của tín hiệu). Điều này làm cho tín hiệu tiếng nói chứa hỗn hợp các tín hiệu thành phần tạo nên từ các cơ quan này, dẫn đến thuật toán tính F0 phải xử lý cả các phần tín hiệu không liên quan đến tính tuần hoàn của tín hiệu.

1.6. Tổng kết chương

Tiếng nói là sóng âm lan truyền trong môi trường không khí. Tiếng nói được tạo ra bởi độ rung của dây thanh trong hệ thống phát âm. Con người thu nhận âm thanh thông qua bộ phận thu nhận âm thanh để xử lý thông tin được truyền đi từ người nói.

Trong xử lý tín hiệu tiếng nói, tiếng nói được biểu diễn trên miền thời gian và trên miền tần số. Tín hiệu tiếng nói được biểu diễn trên miền thời gian là đồ thị biểu diễn tín hiệu tiếng nói theo trục thời gian. Tín hiệu tiếng nói được biểu diễn trên miền tần số là đồ thị biểu diễn tín hiệu tiếng nói theo trục tần số.

Tiếng nói ở mỗi người đều có đặc trưng khác nhau. Các đặc trưng này được tạo nên từ âm sắc, cường độ, trường độ. Ở mỗi người, các đại lượng này là khác nhau nên tiếng nói cảm nhận được là khác nhau. Trong lĩnh vực xử lý tín hiệu tiếng nói, F0 là đặc trưng quan trọng của tín hiệu tiếng nói. Để tìm F0 của tín hiệu tiếng nói, cần dùng đến kỹ thuật xử lý ngắn hạn chia tín hiệu tiếng nói thành nhiều khung nhỏ để xử lý.

Việc tính F0 tự động là một trong các bài toán cơ bản của lĩnh vực xử lý tiếng nói. Đã có nhiều thuật toán được đề xuất để tính giá trị F0 của tín hiệu tiếng nói [7][8]. Mỗi thuật toán có những ưu và nhược điểm khác nhau. Trong luận văn, tôi chọn nghiên cứu và cài đặt thuật toán tìm F0 dùng hàm tự tương quan vì tính đơn giản về lý thuyết và cài đặt thực tế. Thuật toán này đã được thử nghiệm trên tín hiệu tiếng nói [9] cũng như tín hiệu âm nhạc [2] và đã cho thấy hiệu quả của nó.

CHƯƠNG 2: TÍNH TẦN SỐ CƠ BẢN DÙNG HÀM TỰ TƯƠNG QUAN

2.1. Mở đầu

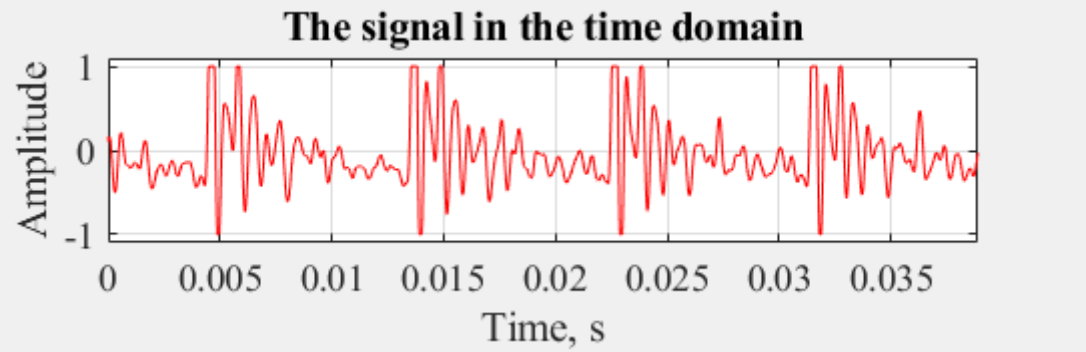
Như đã đề cập trong chương trước, tần số cơ bản (hay F0) của tín hiệu tiếng nói là tham số có ý nghĩa quan trọng trong lĩnh vực xử lý tiếng nói. Tìm được F0 chính xác là tiền đề để tiến hành các nghiên cứu khác trong lĩnh vực này.

Để tính được F0, trong phạm vi của luận văn, tôi nghiên cứu hàm tự tương quan đối với tín hiệu tiếng nói. Trong thực tế, khi nghiên cứu về tín hiệu tuần hoàn, hàm tự tương quan được sử dụng nhiều vì từ hàm này dễ dàng xác định ra được chu kỳ cơ bản T0 của tín hiệu, từ đó suy ra tần số cơ bản F0 là nghịch đảo của T0.

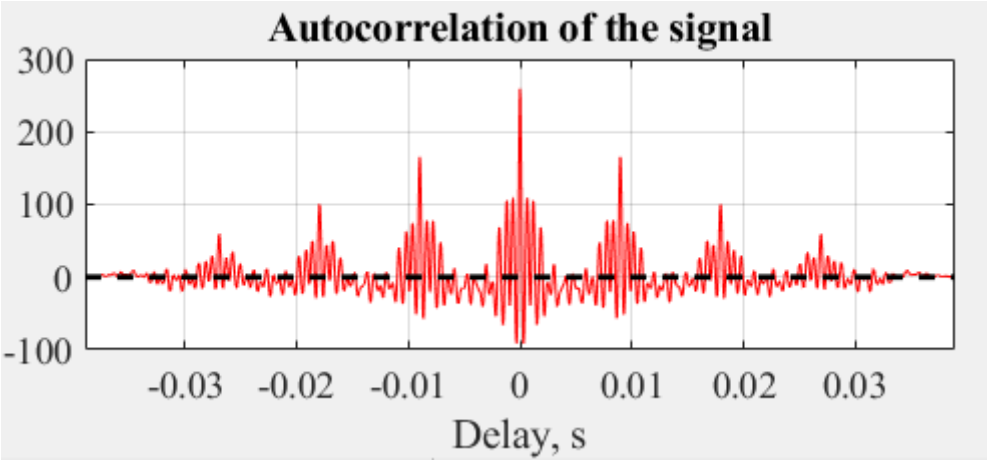
Ngoài ra, chuỗi giá trị F0 sau khi tính được bằng thuật toán tự tương quan có cải tiến trên các khung tín hiệu thường vẫn tồn tại một vài giá trị F0 tăng hoặc giảm đột biến so với các giá trị F0 còn lại. Do đó, cần có thuật toán lọc trung vị để loại bỏ các giá trị đột biến này nhằm thu được đường F0 đủ trơn như mong muốn. Điều này xuất phát từ thực tế là tần số rung của dây thanh của một người không thể biến đổi quá nhiều trong quá trình phát âm do cấu tạo của thanh quản.

2.2. Hàm tự tương quan và ứng dụng để tính F0

Trong xử lý tín hiệu số nói chung và xử lý tín hiệu tiếng nói nói riêng, hàm tự tương quan dùng để biến đổi tín hiệu tuần hoàn thành một tín hiệu tuần hoàn khác có các điểm cực đại có thể xác định được dễ dàng, nhờ đó ứng dụng để xác định chu kỳ cơ bản T0 và tần số cơ bản F0 [5]. Hình 2.1 minh hoạ một ví dụ như vậy.



Hình 2.1 – Một đoạn tín hiệu tuần hoàn trên miền thời gian



Hình 2.2 – Hàm tự tương quan của đoạn tín hiệu tuần hoàn trong Hình 2.1

Hàm tự tương quan của tín hiệu được xác định bởi công thức [4]:

$$r_{xx}(\mathrm{l}) = \lim_{N=-\infty} \frac{1}{(2N+1)} \sum_{n=-N}^{N} x(n)x(n+l) \qquad (2.1)$$

trong đó: $r_{xx}(l)$ là giá trị hàm tự tương quan theo độ trễ l, $(2N+1)$ là độ dài khung tín hiệu, $x(n)$ là biên độ tín hiệu tại thời điểm n.

Hàm tự tương quan có các tính chất sau:

- Là một hàm chẵn: $r_{xx}(l)=r_{xx}(-l)$;
- Đạt giá trị cực đại tại $l=0$: $|r_{xx}(l)| \le r_{xx}(0)$ với mọi l;
- Đại lượng $r_{xx}(0)$ bằng năng lượng của tín hiệu tiếng nói.

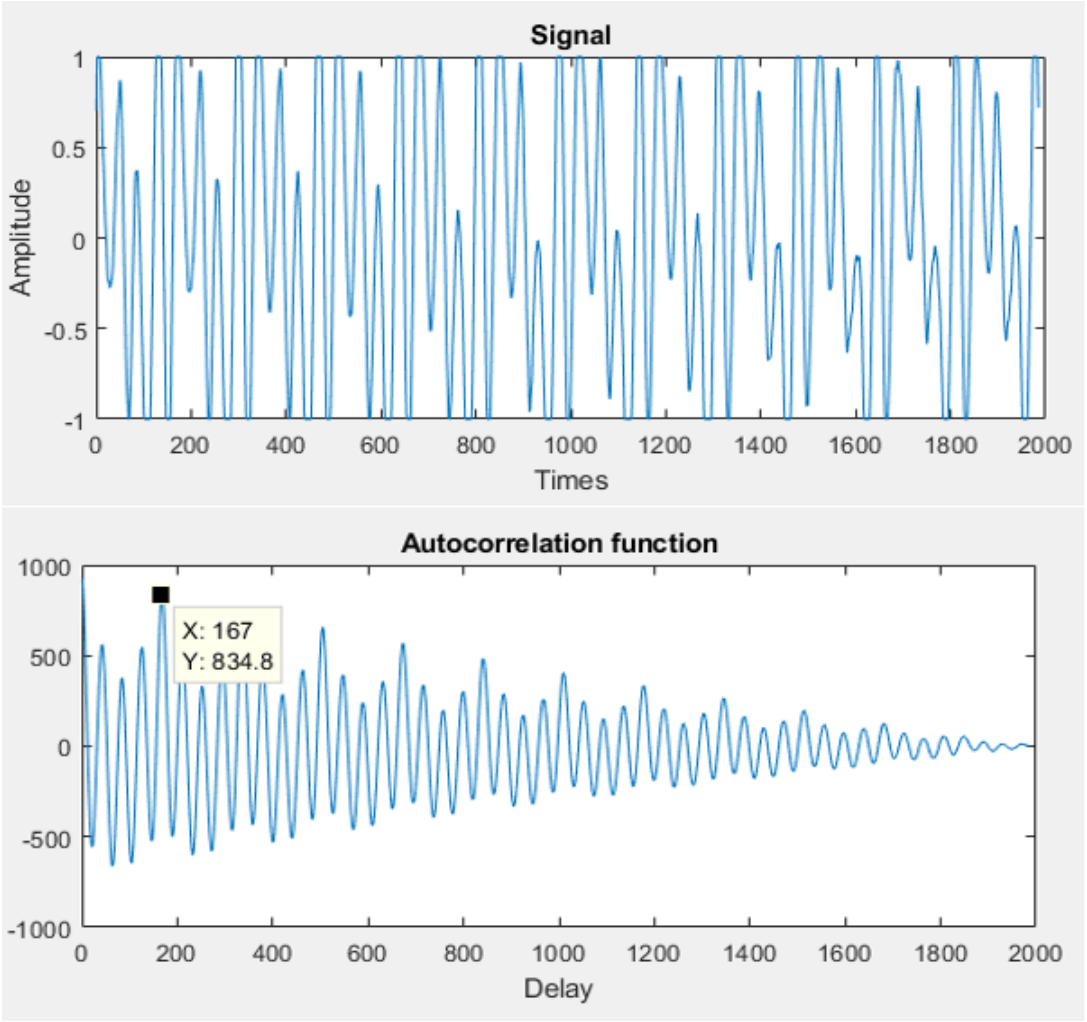
Khi xử lý tín hiệu dùng kỹ thuật xử lý ngắn hạn (phần 1.4), ta chia tín hiệu tiếng nói thành các khung tín hiệu có độ dài hữu hạn và công thức tự tương quan trở thành [2]:

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j x_{j+\tau} \qquad (2.2)$$

trong đó x_j là biên độ tín hiệu tại thời điểm j, $r_t(\tau)$ là giá trị của hàm tự tương quan theo độ trễ τ tại khung tín hiệu t, và W là độ dài của khung tín hiệu.

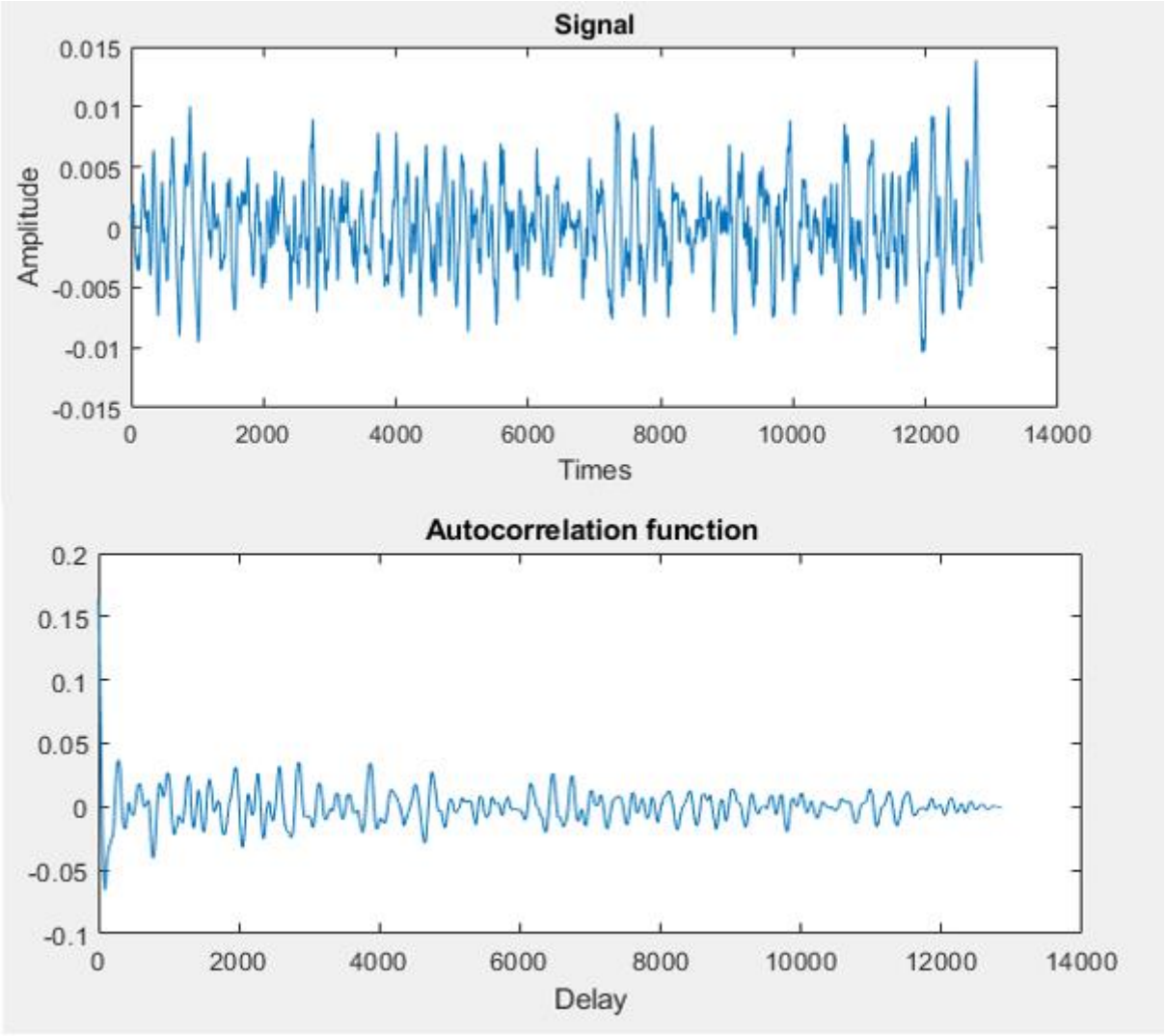
Nếu $T0$ là chu kỳ cơ bản của tín hiệu tuần hoàn, khi đó các giá trị độ trễ: 0, $\pm T0$, $\pm 2T0$,… sẽ là các điểm mà hàm tự tương quan đạt cực đại cục bộ. Đây là ý tưởng chính để xác định F0 của tín hiệu tiếng nói bằng hàm tự tương quan.

Tiếng nói có 2 loại âm: hữu thanh và vô thanh (phần 1.3). Tín hiệu của âm hữu thanh có dạng sóng gần như tuần hoàn nên hàm tự tương quan của nó sẽ xuất hiện các điểm cực đại cục bộ tại các độ trễ có giá trị bằng bội số nguyên lần của chu kỳ cơ bản. Hình 2.3 minh hoạ một ví dụ về đoạn tín hiệu và hàm tự tương quan của một âm hữu thanh có chu kỳ cơ bản T0 = 167 (mẫu), chính là giá trị độ trễ ứng với điểm cực đại cục bộ có biên độ lớn nhất của hàm tự tương quan.



Hình 2.3 – Tín hiệu (trên) và hàm tự tương quan (dưới) của một âm hữu thanh

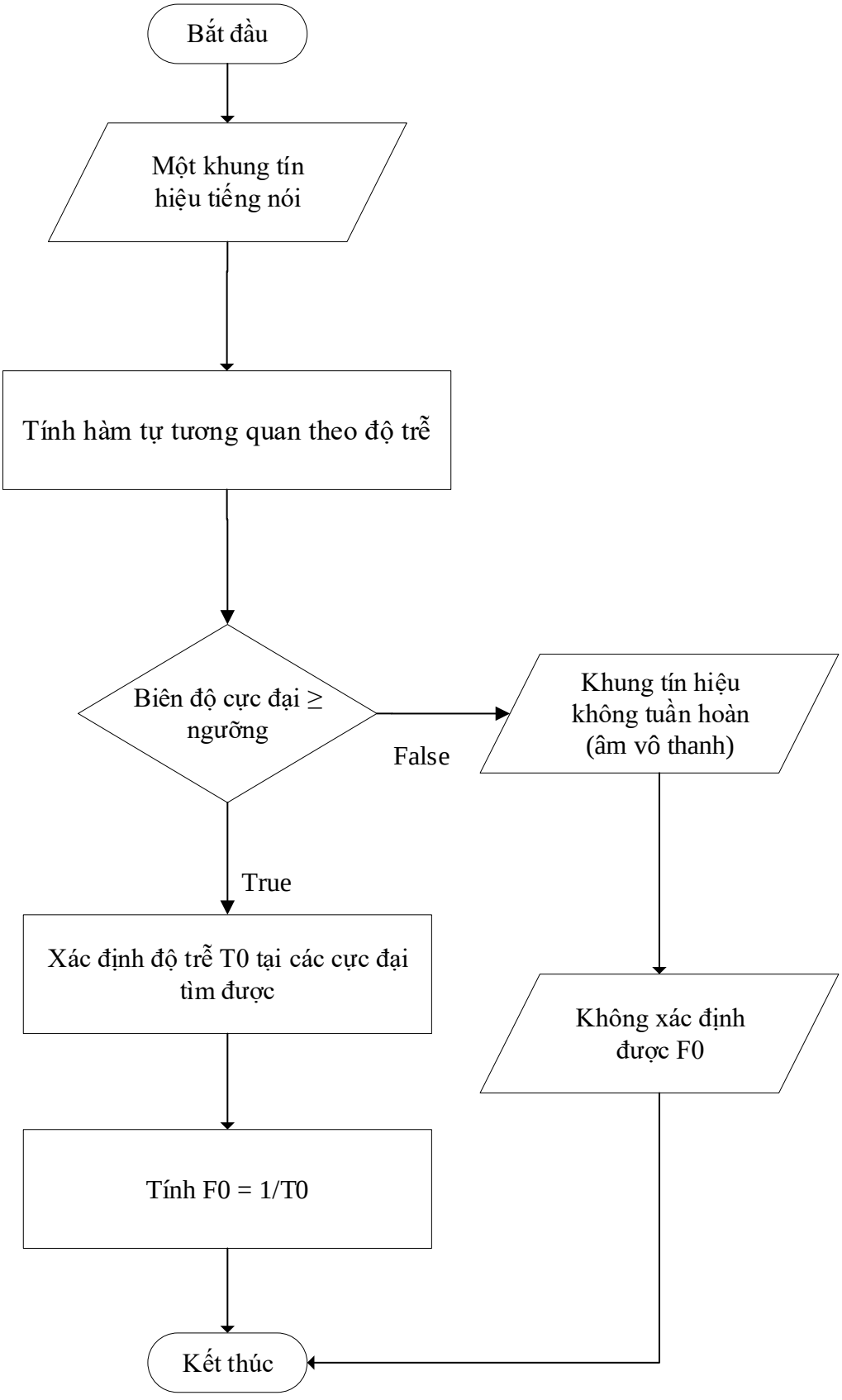
Ngược lại, tín hiệu của âm vô thanh có dạng sóng không tuần hoàn nên hàm tự tương quan của nó sẽ không có tính chất tương tự như âm hữu thanh. Hình 2.4 minh hoạ một ví dụ về đoạn tín hiệu và hàm tự tương quan của một âm vô thanh. Chúng ta khó thấy rõ các điểm cực đại cục bộ của hàm tự tương quan nằm ở đâu, và các điểm cực đại này cũng không nằm cách đều nhau như trường hợp âm hữu thanh mà nằm rải rác một cách ngẫu nhiên. Hai ví dụ trên cho thấy, giá trị cao hay thấp của điểm cực đại cục bộ có biên độ lớn nhất của hàm tự tương quan có thể dùng để phân biệt một khung tín hiệu là hữu thanh hay vô thanh.



Hình 2.4 – Tín hiệu (trên) và hàm tự tương quan (dưới) của một âm vô thanh

2.3. Thuật toán tính F0

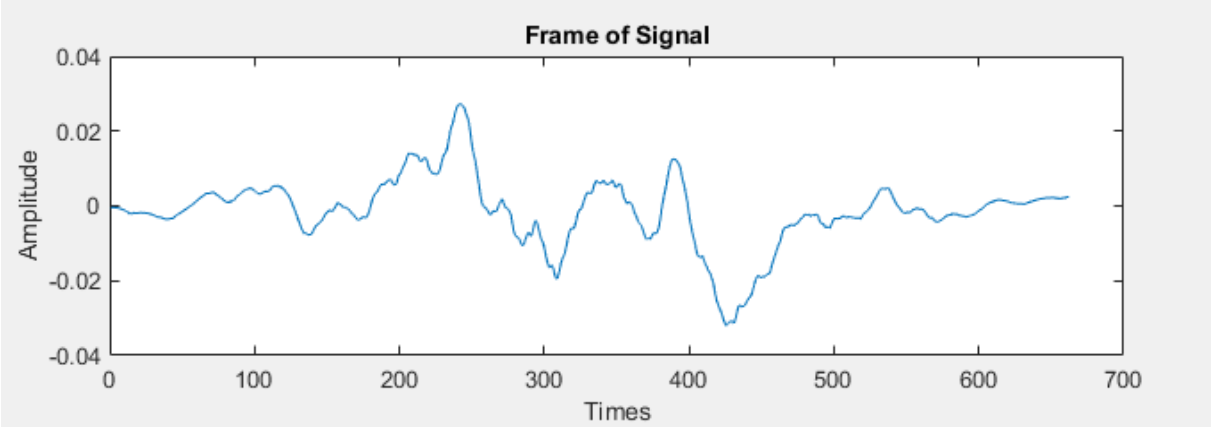
Với những phân tích trong phần 2.2, tôi đưa ra thuật toán tính F0 của một khung tín hiệu dựa trên hàm tự tương quan như Hình 2.5:



Hình 2.5 – Thuật toán tìm F0 dùng hàm tự tương quan

Thuật toán trên được diễn giải như sau. Bằng kỹ thuật xử lý ngắn hạn, tín hiệu tiếng nói đầu vào được chia nhỏ thành các khung tín hiệu ngắn (có độ dài từ 10 ms đến 30 ms) để xử lý. Trong luận văn, tôi thực hiện phân khung bằng hàm cửa sổ Hamming [4]. Hàm cửa sổ Hamming được xác định bởi công thức:

$$w(n) = 0.54 - 0.46\cos(2\pi \frac{n}{N}), \quad 0 \le n \le N \tag{2.3}$$



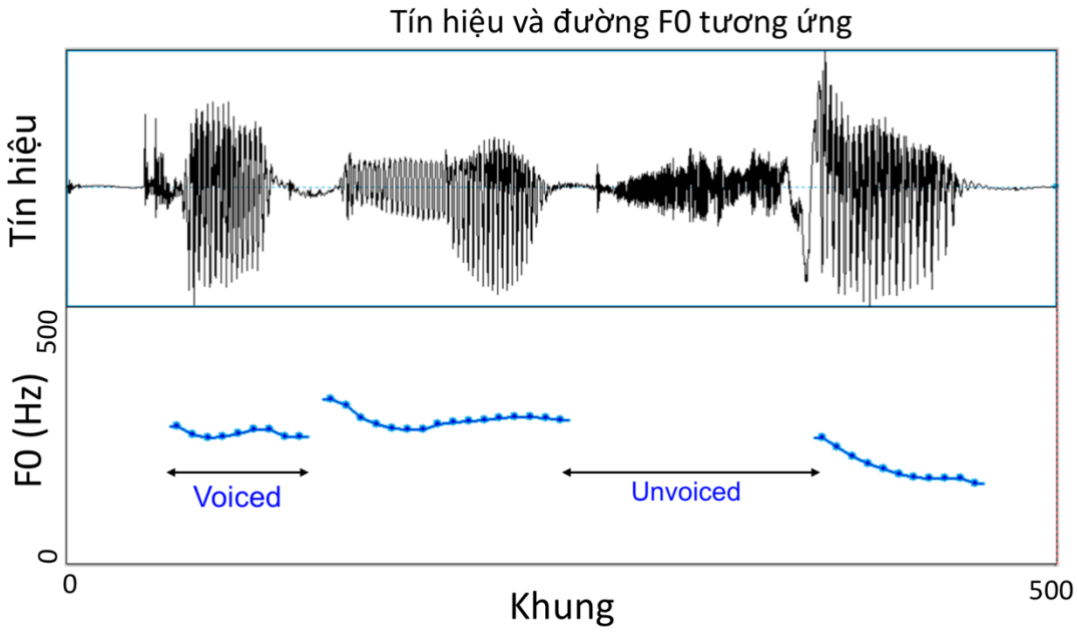
Hình 2.6 – Ví dụ về một khung tín hiệu có độ dài 662 mẫu (tương đương 15 ms với tần số lấy mẫu 44100 Hz).

Sau khi tín hiệu được cắt thành từng khung, mỗi khung tín hiệu cần được phân loại thuộc về âm hữu thanh hoặc âm vô thanh. Nếu khung tín hiệu thuộc về âm vô thanh (nghĩa là khung tín hiệu không tuần hoàn) thì F0 không xác định. Nếu khung tín hiệu thuộc về âm hữu thanh thì sẽ có giá trị F0 xác định. Việc phân loại hữu thanh/vô thanh dựa trên hàm tự tương quan của khung tín hiệu như sau: thuật toán cần xác định điểm cực đại cục bộ có biên độ lớn nhất của hàm tự tương quan (trừ điểm cực đại toàn cục tại vị trí độ trễ τ=0). Nếu điểm cực đại cục bộ có biên độ lớn nhất tìm được có biên độ nhỏ hơn một ngưỡng nào đó (thường là 30% giá trị biên độ của điểm cực đại toàn cục [4]) thì đó là âm vô thanh, ngược lại là âm hữu thanh.

Nếu khung tín hiệu thuộc về âm hữu thanh thì giá trị độ trễ tại điểm cực đại có biên độ lớn nhất của hàm tự tương quan chính là chu kỳ cơ bản T0 của khung tín hiệu. Từ đó ta xác định được F0 của khung tín hiệu đang xét theo công thức:

$$F0 = \frac{1}{T0} \tag{2.4}$$

Việc chia tín hiệu thành chuỗi các khung để xử lý và tính F0 dẫn đến kết quả đường F0 thu được có dạng như Hình 2.7 với giá trị F0 xác định tại các khung hữu thanh (voiced frames) và F0 không xác định tại các khung vô thanh (unvoiced frames).



Hình 2.7 – Ví dụ minh hoạ tín hiệu và kết quả tính F0 của nó.

2.4. Các tham số quan trọng của thuật toán

Phần này trình bày 2 tham số quan trọng ảnh hưởng nhiều đến độ chính xác của thuật toán tự tương quan, đó là độ dài khung tín hiệu và ngưỡng xác định hữu thanh/vô thanh. Các tham số này sẽ được khảo sát thực nghiệm trong phần 3.6 và 3.8.

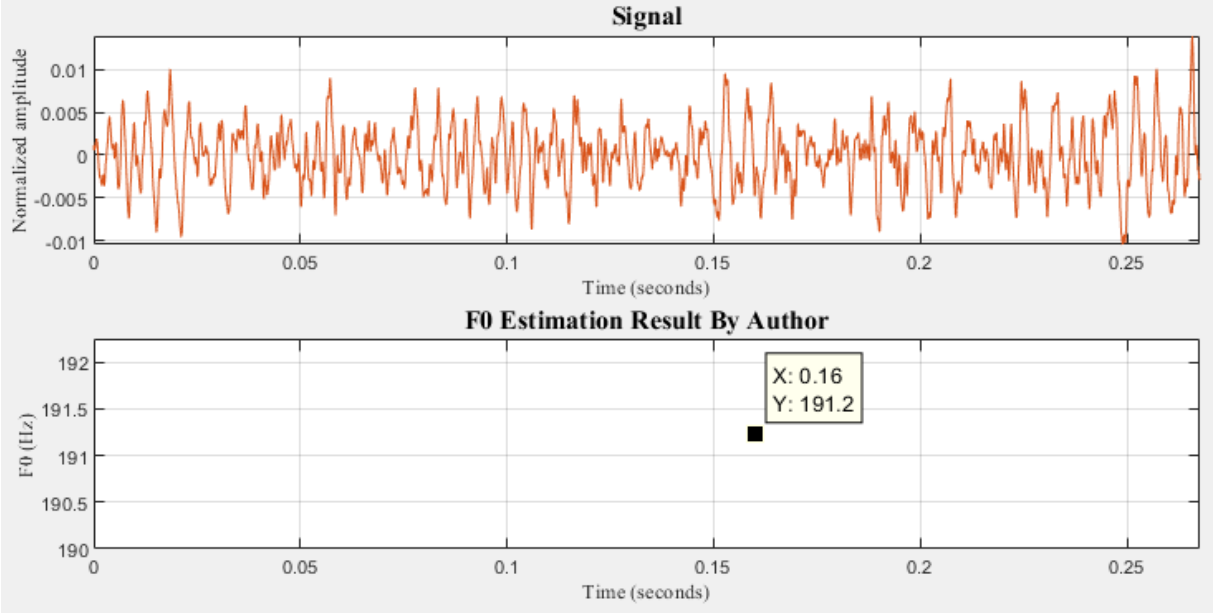
2.4.1. Độ dài khung tín hiệu

Thuật toán tính F0 dựa trên hàm tự tương quan có sử dụng kỹ thuật xử lý ngắn hạn (phân tín hiệu thành nhiều khung nhỏ) nên việc chọn loại cửa sổ và độ dài cửa sổ thích hợp là quan trọng. Có rất nhiều hàm cửa sổ có thể được sử dụng trong kỹ thuật xử lý tín hiệu ngắn hạn: Hamming, Hanning, Blackman, tam giác, chữ nhật [5]. Trong luận văn, tôi chọn cửa sổ Hamming do tính phổ dụng của nó trong xử lý tín hiệu tiếng nói [4]. Về độ dài cửa sổ, một cửa sổ có độ dài từ 10 ms đến 30 ms (để đảm bảo các tính chất của tín hiệu tiếng nói tương đối ổn định trong khung tín hiệu) và bao gồm ít nhất 2 chu kỳ liên tiếp của tín hiệu là điều kiện cần để xác định được chu kỳ cơ bản T0, từ đó suy ra F0, của tín hiệu. Tuy nhiên, nếu một cửa sổ chứa quá nhiều chu kỳ tín hiệu lại có thể làm cho thuật toán dễ mắc lỗi cao độ ảo [2], trong đó giá trị F0 tìm được thường gấp đôi (hoặc gấp ba) hay chỉ bằng 1/2 (hoặc 1/3) giá trị F0 thực sự.

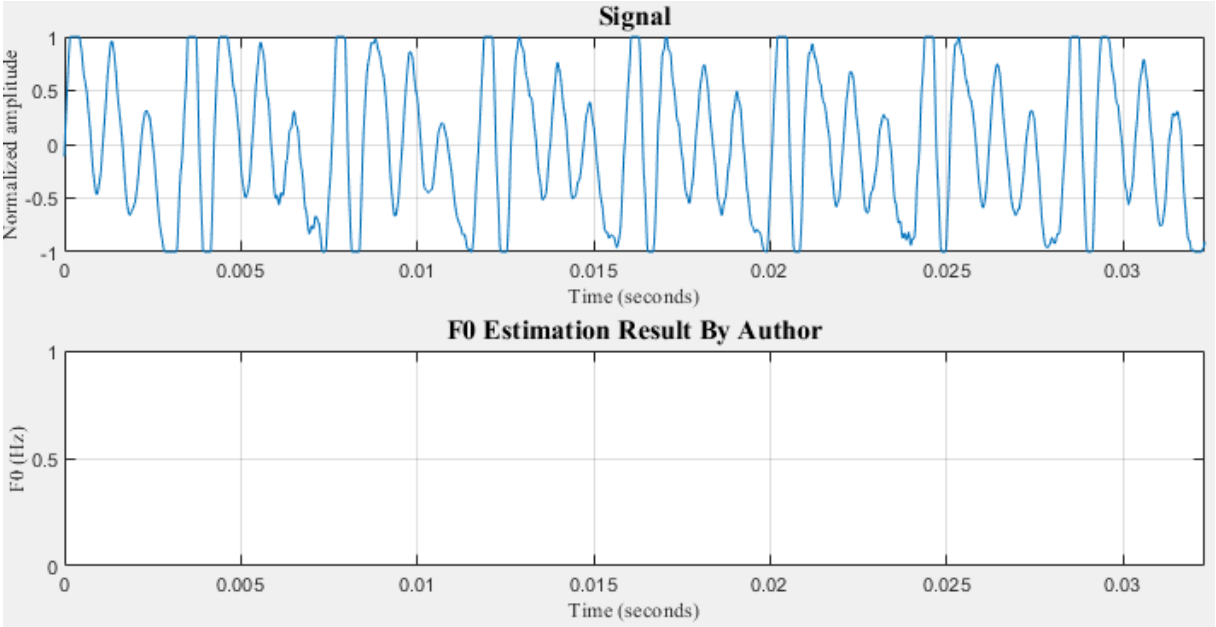
2.4.2. Ngưỡng xác định hữu thanh/vô thanh

Trong thuật toán tìm F0 dùng hàm tự tương quan, có một tham số quan trọng nữa đó là ngưỡng để xác định một khung tín hiệu là của âm hữu thanh hay của âm vô thanh.

Việc tăng hoặc giảm ngưỡng này ảnh hưởng đến việc xác định âm hữu thanh hoặc âm vô thanh của đoạn tín hiệu tiếng nói. Nếu ngưỡng này đặt ra là quá thấp, khi tính F0, các khung tín hiệu vô thanh sẽ bị nhầm thành các khung tín hiệu hữu thanh. Nếu ngưỡng này đặt ra là quá cao, khi tính F0, các khung tín hiệu hữu thanh sẽ bị nhầm thành các khung tín hiệu vô thanh.



Hình 2.8 - Tín hiệu của âm vô thanh bị xác định nhầm thành âm hữu thanh, dẫn đến xác định được F0 = 191,2 Hz tại 0,16 giây



Hình 2.9 - Tín hiệu của âm hữu thanh bị xác định nhầm thành âm vô thanh và không xác định được giá trị F0 nào

2.5. Lọc trung vị

2.5.1. Cơ sở lý thuyết

Trong hầu hết các ứng dụng xử lý tín hiệu, làm trơn tuyến tính hầu như được sử dụng để loại bỏ các thành phần nhiễu trong tín hiệu. Tuy nhiên, với một vài ứng dụng xử lý tiếng nói, làm trơn tuyến tính không hoạt động hiệu quả do tính chất của tín hiệu được làm trơn. Một ví dụ là đường F0 xác định từ tín hiệu tiếng nói không những chứa các giá trị thay đổi một cách bất thường (outliers) so với các giá trị lân cận (Hình 2.10) mà còn gián đoạn tại các vùng chuyển tiếp giữa âm vô thanh và âm hữu thanh (Hình 2.7). Một bộ lọc tuyến tính thông thấp sẽ không kéo được giá trị F0 bất thường về gần các giá trị đúng và còn làm méo đường F0 tại các điểm gián đoạn. Trong trường hợp này, một kỹ thuật làm trơn phi tuyến như lọc trung vị là cần thiết.

Làm trơn trung vị (median smoothing) là kỹ thuật lọc phi tuyến được sử dụng phổ biến trong xử lý tín hiệu. Nó có ưu điểm là loại bỏ được giá trị nhảy vọt so với các giá trị lân cận mà vẫn bảo toàn các điểm gián đoạn trong tín hiệu. Giá trị đầu ra của bộ lọc trung vị ứng với giá trị đầu vào *x(n)*, ký hiệu là $M_N[x(n)]$, là giá trị trung vị (median) của *N* giá trị *x(n-L), …, x(n),..., x(n+L) (với N=2L+1 là số nguyên dương lẻ)*. Các giá trị trung vị với chiều dài cửa sổ lọc *N* có các tính chất sau [4]:

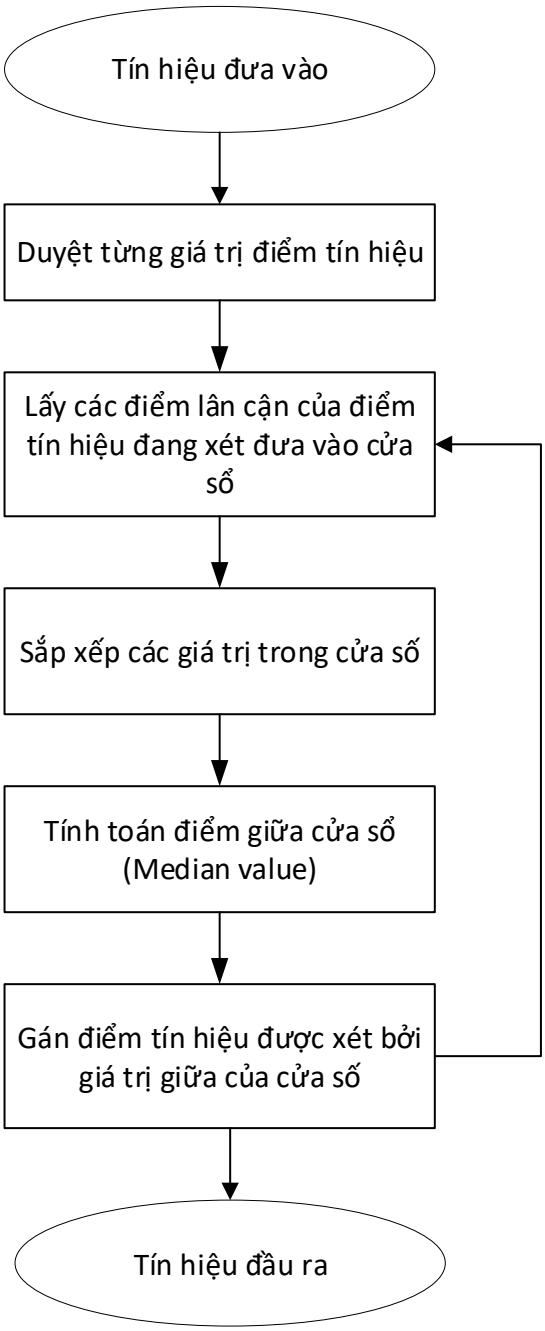
- $M_N[\alpha x(n)] = \alpha M_N[x(n)]$;

- Các giá trị trung vị không làm nhoè các điểm gián đoạn trong tín hiệu nếu tín hiệu không có điểm gián đoạn nào khác trong phạm vi *N*/2 mẫu (nghĩa là các điểm gián đoạn phải cách nhau đủ xa);

- Các giá trị trung vị bám theo, một cách gần đúng, xu hướng có dạng đa thức bậc thấp của tín hiệu.

Mặc dù lọc trung vị giữ lại được các gián đoạn sắc nét trong tín hiệu, kỹ thuật này lại thường không loại bỏ hoàn toàn được các thành phần giống nhiễu của tín hiệu. Khi đó, người ta kết hợp lọc trung vị với lọc thông thấp tuyến tính để tận dụng ưu điểm của cả 2 phương pháp.

2.5.2. Thuật toán lọc trung vị

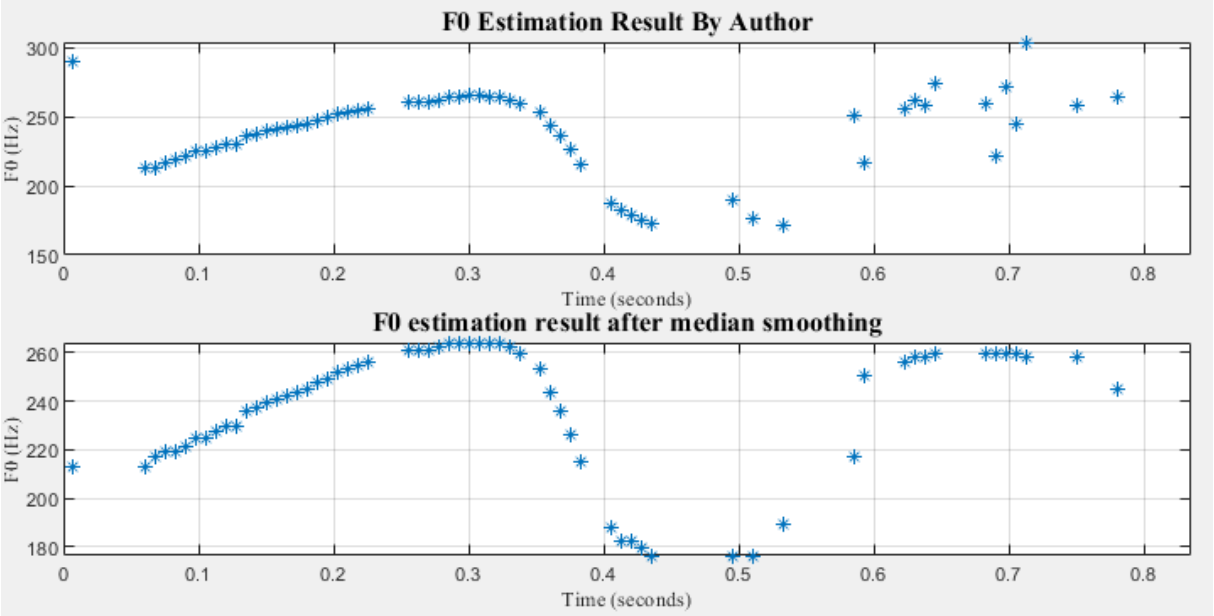


Hình 2.10 – Sơ đồ khối thuật toán lọc trung vị

Trong luận văn, tín hiệu đưa vào thuật toán lọc trung vị là chuỗi giá trị F0 của tín hiệu tiếng nói được xác định bởi thuật toán tự tương quan. Việc lọc trung vị sẽ áp dụng tuần tự cho từng điểm từ đầu đến cuối tín hiệu đầu vào. Với kích thước N được định sẵn của cửa sổ lọc trung vị, N giá trị lân cận 2 bên của điểm tín hiệu đang xét sẽ được điền vào cửa sổ. Sau khi tính toán, giá trị trung vị của cửa sổ này sẽ được gán cho điểm tín hiệu đang xét. Thuật toán sẽ được lặp lại cho đến khi kết thúc tín hiệu đầu vào.

Đối với các giá trị F0 ở gần hai biên thì cửa sổ lọc trung vị sẽ bị thiếu giá trị do không có đủ N giá trị lân cận 2 bên. Để khắc phục điều này, trước tiên tôi bổ sung N/2 giá trị

F0 ở biên trái vào trước chuỗi giá trị F0 và bổ sung N/2 giá trị F0 ở biên phải vào sau chuỗi giá trị F0, sau đó tôi mới tiến hành chạy lọc trung vị trên chuỗi giá trị F0 ban đầu.



Hình 2.11 – Đường F0 trước (hình trên) và sau khi lọc trung vị (hình dưới)

2.5.3. Kích thước bộ lọc

Độ dài N của cửa sổ lọc trung vị (còn gọi là kích thước bộ lọc) là tham số quan trọng để bộ lọc trung vị có thể hoạt động đúng. Nếu kích thước càng lớn, giá trị tính toán sẽ được thu hẹp và điểm lỗi sẽ gần hơn so với các điểm trơn hoặc các điểm đúng còn lại. Tuy nhiên, kích thước quá lớn của bộ lọc cũng sẽ ảnh hưởng đến tốc độ tính toán cũng như điểm bất thường được sửa lỗi. Nếu trong tín hiệu có quá nhiều điểm bất thường thì điểm bất thường cần sửa lỗi không có tác dụng khi sử dụng bộ lọc trung vị. Do đó, kích thước của cửa sổ lọc cần chọn phải phù hợp với tín hiệu. Tham số này được khảo sát thực nghiệm trong phần 3.5.

2.6. Tổng kết chương

Về bản chất, hàm tự tương quan là hàm biến đổi tín hiệu từ miền thời gian sang miền độ trễ. Do đó, để tính được F0, cần phải có các kỹ thuật khác liên quan đến miền độ trễ được áp dụng trong thuật toán tìm F0 của tín hiệu tiếng nói. Qua đó sẽ làm cho giá trị F0 tìm được chính xác hơn.

Để đánh giá được thuật toán và hàm tự tương quan trong việc tính F0 đối với tín hiệu tiếng nói thu được, trong chương 3 tôi sẽ trình bày về ứng dụng sử dụng hàm tự tương

quan cũng như đánh giá hàm tự tương quan phát triển được so với cách thủ công để tính F0.

CHƯƠNG 3: TRIỂN KHAI VÀ ĐÁNH GIÁ THUẬT TOÁN

3.1. Mở đầu

Trong chương này, tôi tiến hành cài đặt thuật toán tính F0 dùng hàm tự tương quan trên Matlab [6]. Đồng thời, tôi dùng thuật toán lọc trung vị để làm trơn kết quả tính F0 nhận được từ thuật toán tự tương quan.

Để rút ra được kết quả và nhận xét hàm tự tương quan khi thực hiện tính F0 của tín hiệu tiếng nói, tôi cài đặt hai hàm tự tương quan, một hàm là do tôi tự triển khai, và một hàm thư viện của Matlab (hàm **xcorr()**). Đồng thời, tôi so sánh kết quả tính F0 tự động bởi thuật toán với cách đo thủ công.

Để đánh giá độ chính xác của một thuật toán tính F0, người ta thường dùng 2 thước đo gồm: lỗi xác định hữu thanh/vô thanh và sai số của giá trị F0 [8]. Lỗi xác định hữu thanh/vô thanh lại được chia thành 2 loại lỗi sau:

- Lỗi nhầm hữu thanh thành vô thanh: là tỷ lệ lỗi khung tín hiệu tuần hoàn (ứng với âm hữu thanh) bị xác định nhầm thành khung không tuần hoàn (ứng với âm vô thanh).
- Lỗi nhầm vô thanh thành hữu thanh: là tỷ lệ lỗi khung tín hiệu không tuần hoàn (ứng với âm vô thanh) bị xác định nhầm thành khung tuần hoàn (ứng với âm hữu thanh).

Sai số của giá trị F0 chỉ được tính trên các khung tín hiệu được xác định thuộc âm hữu thanh, dựa trên các giá trị F0 chuẩn (thường đo bằng phương pháp thủ công) và các giá trị F0 tính tự động bởi thuật toán.

Tuy nhiên, các thước đo trên chỉ có thể đánh giá được trên tập tín hiệu đã xác định trước mỗi khung tín hiệu là hữu thanh hay vô thanh, và nếu là khung hữu thanh thì F0 có giá trị chuẩn bằng bao nhiêu. Do không đủ thời gian để xây dựng tập dữ liệu F0 chuẩn cho tín hiệu tiếng nói bất kỳ (ví dụ như của cả câu), trong luận văn tôi chỉ khảo sát tín hiệu của các nguyên âm vì mỗi nguyên âm đều là âm hữu thanh và giá trị F0 của người nói có thể coi là thay đổi không đáng kể trong thời gian phát âm. Khi đó, sai số của thuật toán tính F0 trên mỗi tín hiệu nguyên âm được tính bằng độ lệch tuyệt đối giữa giá trị F0 chuẩn được đo thủ công và giá trị F0 tự động tính bởi thuật toán.

3.2. Môi trường phát triển

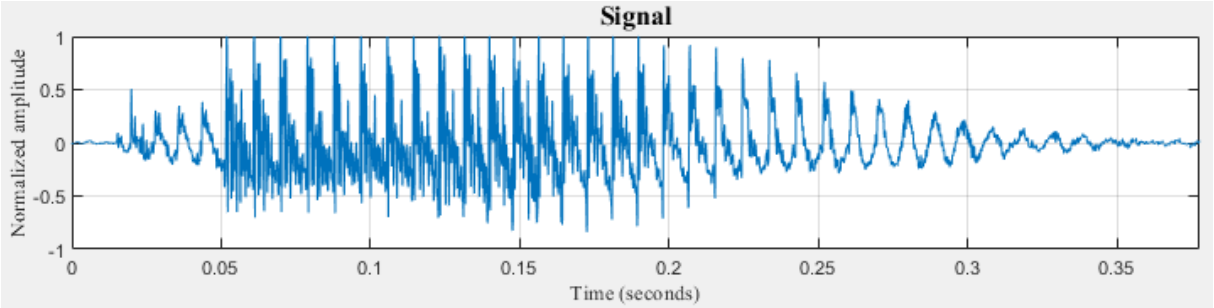
Tôi cài đặt thuật toán và tiến hành thực nghiệm trên máy tính có cấu hình:

- Hệ điều hành: Windows 10 Ultimate x64
- Bộ nhớ trong: 8GB
- Bộ vi xử lý: Intel® Core™ i5-6260U CPU @ 1.80GHz

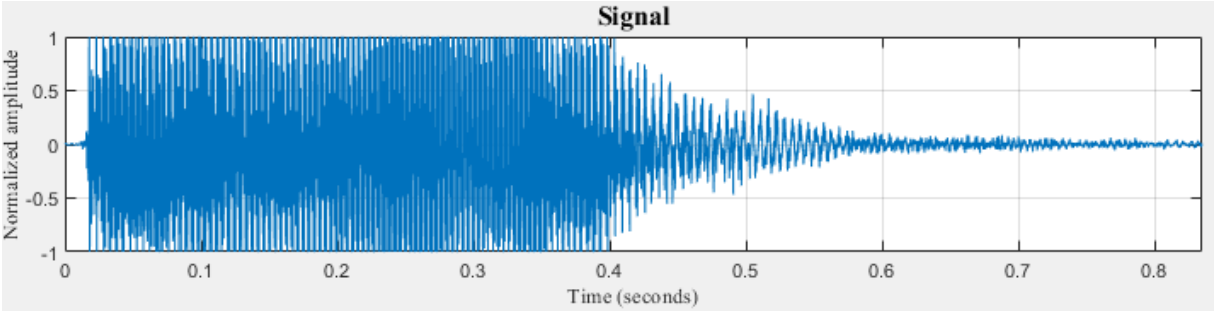
Phần mềm được sử dụng: Matlab – Phiên bản R2018a

3.3. Dữ liệu thử nghiệm

Việc khảo sát tín hiệu của nhiều âm nói bởi nhiều người khác nhau là cần thiết để đánh giá hiệu quả của thuật toán. Tôi đã thu thập tín hiệu tiếng nói của năm nguyên âm /a/, /e/, /i/, /o/, /u/ trong điều kiện phòng với ba giọng nam và ba giọng nữ của người trưởng thành. Các tín hiệu được thu ở tần số lấy mẫu 44100 Hz, đơn kênh (mono), và lưu trong các file .wav theo định dạng PCM của Microsoft.



Hình 3.1 – Tín hiệu nguyên âm /a/ của một người nam

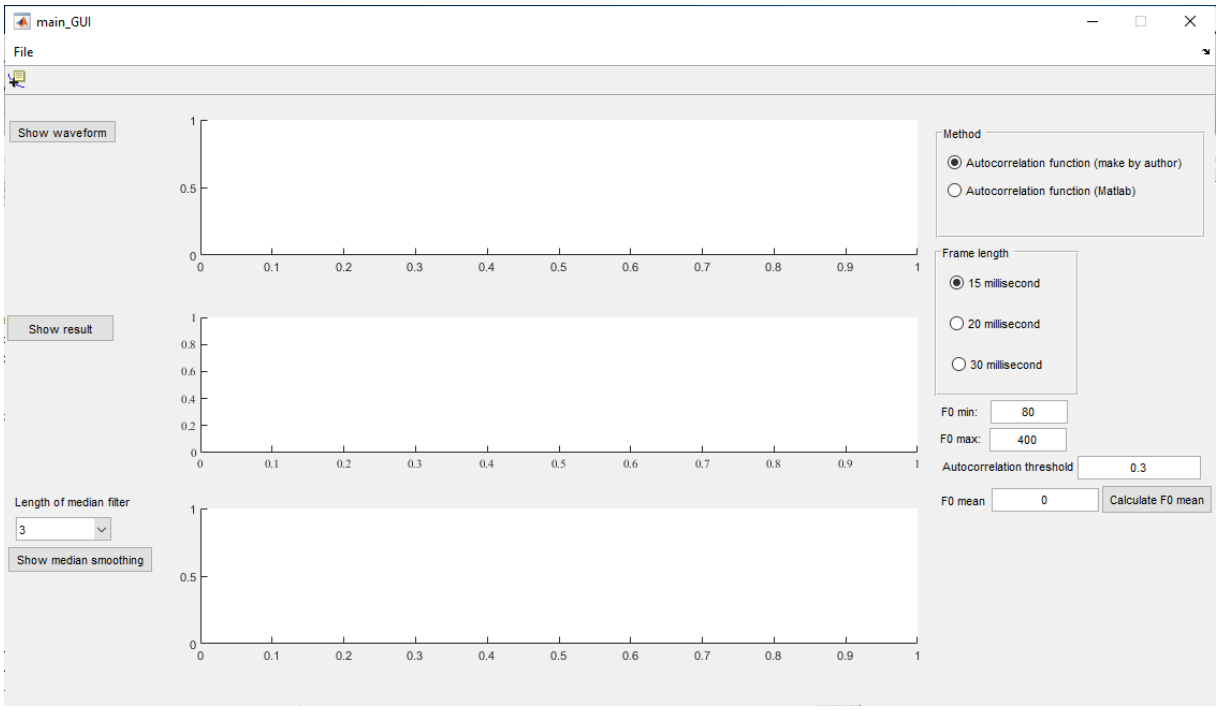


Hình 3.2 – Tín hiệu nguyên âm /a/ của một người nữ

3.4. Demo ứng dụng

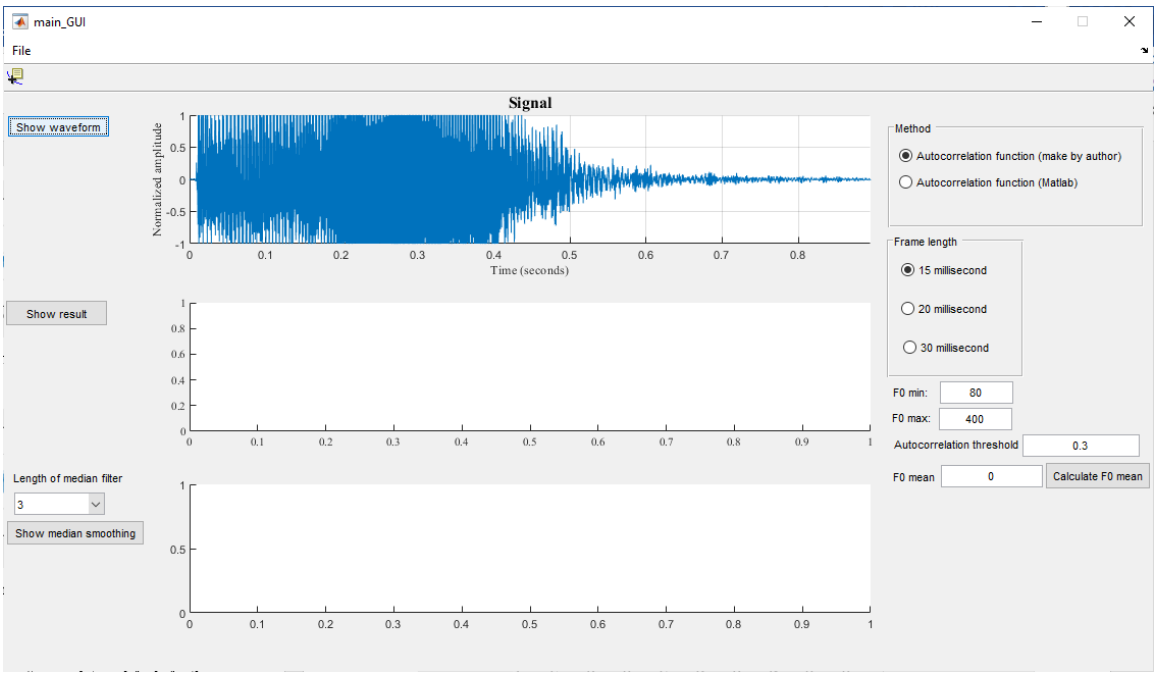
Ứng dụng được viết trên phần mềm Matlab với ba chức năng cơ bản: phần hiển thị sóng âm của tín hiệu tiếng nói, phần hiển thị kết quả tính F0 của hàm tự tương quan tự

lập trình và hàm tự tương quan của Matlab, và phần cuối cùng là kết quả của hàm lọc trung vị trên dữ liệu F0 thu được.



Hình 3.3 – Giao diện chính của chương trình

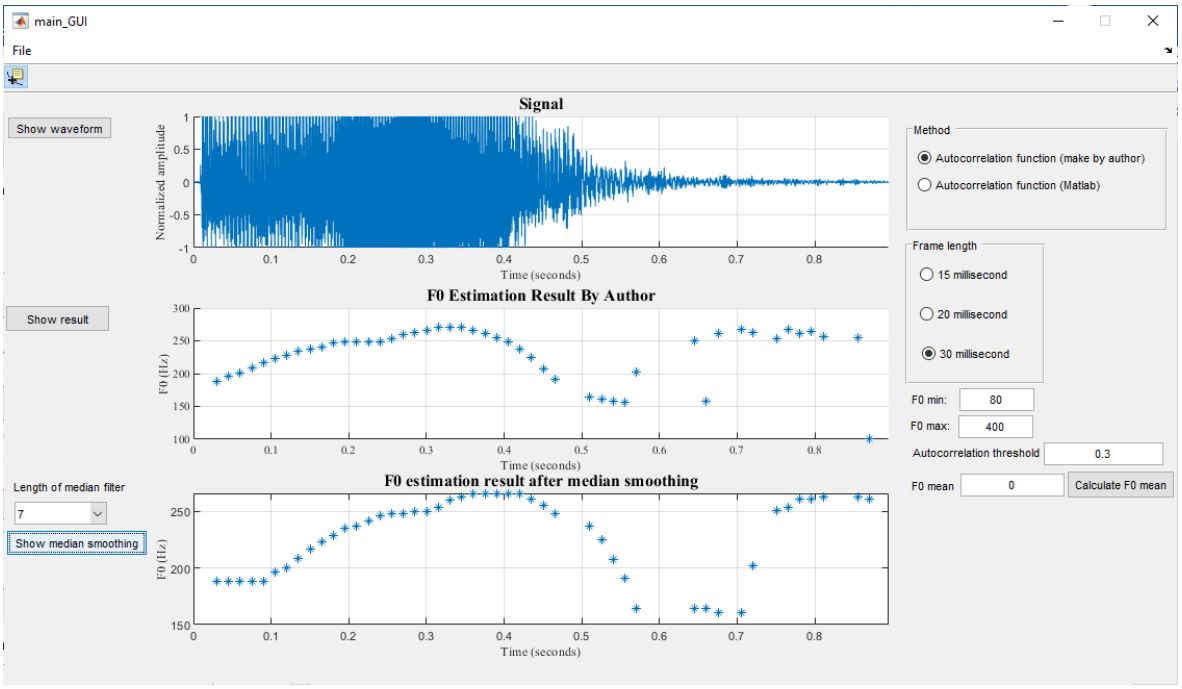
Để tính giá trị F0 của tín hiệu tiếng nói, đầu tiên cần phải mở file bằng cách vào menu File → Open file, sau đó chọn file tín hiệu tiếng nói. Để hiển thị dạng sóng âm của tiếng nói, click vào nút "Show waveform".



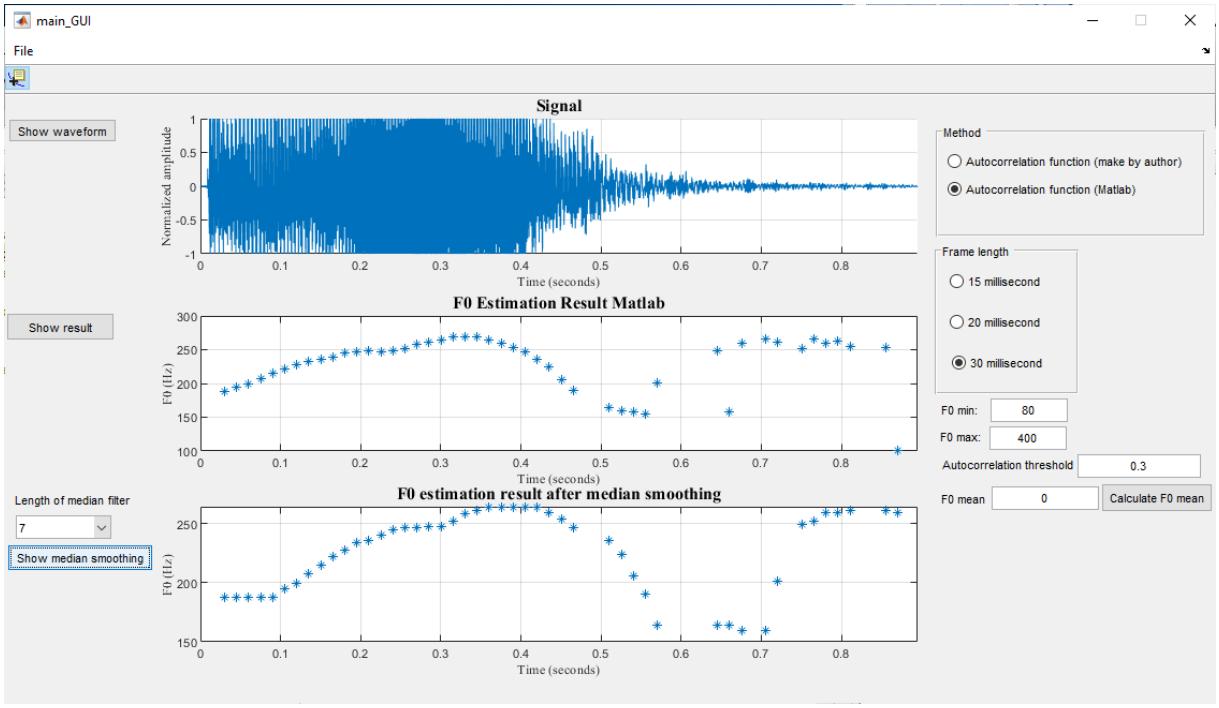
Hình 3.4 – Hiển thị sóng âm của tín hiệu tiếng nói

Để hiển thị kết quả tính toán ứng với hàm tự tương quan của tác giả hoặc hàm tự tương quan của Matlab, click chọn từng radio button tương ứng "Autocorrelation function (make by author)" hoặc "Autocorrelation function (Matlab)" và sau đó click "Show result" để hiển thị kết quả.

Để hiển thị kết quả sau khi lọc trung vị của tính hiệu tiếng nói, click vào nút "Show median smoothing".

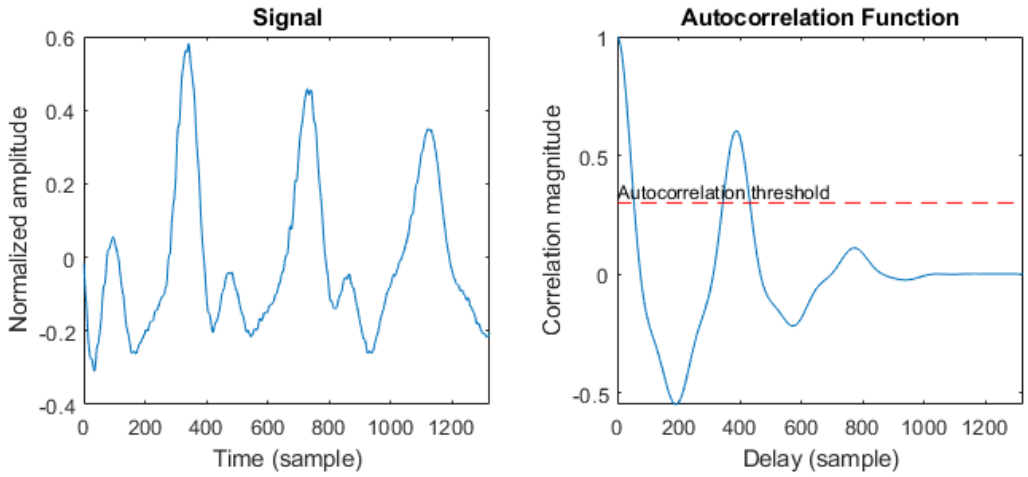


Hình 3.5 – Kết quả tính F0 bằng hàm tự tương quan tự cài đặt và lọc trung vị



Hình 3.6 - Kết quả tính F0 bằng hàm tự tương quan của Matlab và lọc trung vị

Ngoài ra, ứng dụng còn có một chức năng khác đó là chức năng cho phép hiển thị đoạn tín hiệu và kết quả xử lý của đoạn tín hiệu bằng hàm tự tương quan ứng với độ dài khung đã nhập trong ứng dụng. Để thực hiện được chức năng này, cần phải tắt chế độ "data cursor mode" của Matlab từ  sang  , file âm thanh phải được mở. Click vào nút "Show wave form" để hiển thị đồ thị của sóng âm. Từ đồ thị, click chuột trái trên đồ thị để kích hoạt chức năng.



Hình 3.7 – Chức năng xem khung tín hiệu và hàm tự tương quan của khung

3.5. Khảo sát giá trị kích thước bộ lọc trung vị

Trong luận văn, để tính kết quả chính xác hơn, tôi sử dụng hàm lọc trung vị để tính kết quả F0 sau khi được tính bởi hàm tự tương quan. Để xác định được kích thước bộ lọc N có độ tin cậy cao trong việc tính F0 của tín hiệu tiếng nói, tôi tiến hành khảo sát với N lần lượt có giá trị là 3, 5, và 7. Các kích thước của bộ lọc sẽ được khảo sát trên tín hiệu tiếng nói của một người nam và một người nữ ở các âm /a/, /e/, /i/, /o/, /u/. Để khảo sát mang tính đầy đủ hơn, tôi cũng tiến hành khảo sát ở độ dài khung tín hiệu là 15 ms, 20 ms, và 30 ms.

Kết quả thu được ở độ dài khung 15 ms như sau:

Đơn vị đo: Hz

Tín hiệu	**Đo thủ công**	**F0 dùng hàm tự tương quan**		**F0 dùng hàm tự tương quan qua bộ lọc trung vị**					
		F0	**Độ lệch**	**N=3**	**Độ lệch**	**N=5**	**Độ lệch**	**N=7**	**Độ lệch**
/a/	112,0	296,0	184,0	N/A	N/A	N/A	N/A	N/A	N/A

/e/	107,7	230,9	123,2	N/A	N/A	N/A	N/A	N/A	N/A
/i/	117,2	187,2	70,0	N/A	N/A	N/A	N/A	N/A	N/A
/o/	103,2	110,0	6,8	N/A	N/A	N/A	N/A	N/A	N/A
/u/	115,4	337,5	222,1	336,2	220,8	329,7	214,3	N/A	N/A

Bảng 3.1 – Khảo sát kích thước bộ lọc trung vị với một người nam ở khung tín hiệu 15 ms

Đơn vị đo: Hz

Tín hiệu	**Đo thủ công**	**F0 dùng hàm tự tương quan**		**F0 dùng hàm tự tương quan qua bộ lọc trung vị**					
		F0	**Độ lệch**	**N=3**	**Độ lệch**	**N=5**	**Độ lệch**	**N=7**	**Độ lệch**
/a/	315,1	325,0	9,9	325,0	10,0	325,1	10,0	325,0	10,0
/e/	310,7	322,6	12,0	321,9	11,3	322,2	11,5	322,2	11,6
/i/	334,1	333,7	0,5	333,7	0,4	333,6	0,5	333,6	0,5
/o/	317,4	320,8	3,4	320,7	3,4	320,7	3,4	320,7	3,3
/u/	336,7	332,0	4,7	332,0	4,7	331,9	4,8	331,9	4,8

Bảng 3.2 - Khảo sát kích thước bộ lọc trung vị với một người nữ ở khung tín hiệu 15 ms

Ở độ dài khung là 15ms, đối với tín hiệu của người nam thu được, các kết quả hầu hết vẫn chưa có giá trị để thực hiện đánh giá. Ở bảng 3.7, kết quả đo được ở âm /u/, do đó ở độ dài khung 15ms, kết quả đo F0 không đáng tin cây. Nhưng ngược lại, ở độ dài khung này, ở bảng 3.8, kết quả ở tín hiệu nữ thu được lại có giá trị và độ lệch cao nhất là 11,6 Hz. Trong hầu hết các kết quả đo được, với N = 3 cho thấy kết quả tốt hơn so với N = 5 và N = 7. Với N = 5 hoặc N = 7, kết quả thu được không rõ ràng để quyết định kích thước nào của cửa sổ là tốt hơn.

Với độ dài khung 20ms, kết quả thu được như sau:

Đơn vị đo: Hz

Tín hiệu	**Đo thủ công**	**F0 dùng hàm tự tương quan**		**F0 dùng hàm tự tương quan qua bộ lọc trung vị**					
		F0	**Độ lệch**	**N=3**	**Độ lệch**	**N=5**	**Độ lệch**	**N=7**	**Độ lệch**
/a/	112,0	116,7	4,7	116,1	4,1	116,0	4,1	115,2	3,3
/e/	107,7	138,8	31,2	115,1	7,4	114,8	7,1	114,5	6,8
/i/	117,2	123,7	6,5	119,4	2,1	118,6	1,4	118,6	1,3
/o/	103,2	119,9	16,7	114,0	10,8	113,2	10,0	113,1	9,8
/u/	115,4	127,7	12,3	123,1	7,8	123,1	7,7	123,0	7,6

Bảng 3.3 - Khảo sát kích thước bộ lọc trung vị với một người nam

ở khung tín hiệu 20 ms

Đơn vị đo: Hz

Tín hiệu	Đo thủ công	F0 dùng hàm tự tương quan		F0 dùng hàm tự tương quan qua bộ lọc trung vị					
		F0	Độ lệch	N=3	Độ lệch	N=5	Độ lệch	N=7	Độ lệch
/a/	315,1	323,2	8,1	323,1	8,0	323,1	8,1	323,1	8,1
/e/	310,7	321,8	11,2	321,9	11,2	321,8	11,1	321,8	11,1
/i/	334,1	333,9	0,2	333,9	0,2	333,9	0,2	334,0	0,1
/o/	317,4	319,7	2,3	319,4	2,1	319,4	2,0	319,5	2,1
/u/	336,7	321,3	15,4	331,2	5,5	331,2	5,5	331,2	5,5

Bảng 3.4 - Khảo sát kích thước bộ lọc trung vị với một người nữ

ở khung tín hiệu 20 ms

Với độ dài khung tín hiệu là 20 ms, cho thấy với N = 7, tín hiệu hàm tự tương quan qua bộ lọc trung vị đối với giọng nam thu được là tốt nhất. Điều này đã thể hiện rõ qua bảng 3.9 đối với giọng nam thu được. Có một số trường hợp N = 5 có độ lệch bằng N = 7. Tuy nhiên, ở kết quả khác, N = 7 lại cho kết quả tốt hơn so với N = 5.

Đối với giọng nữ, việc chênh lệch giữa các kích thước cửa sổ của bộ lọc trung vị là không nhiều. Nên trong trường hợp này, không thể đánh giá được kích thước của bộ lọc trung vị nào là tốt.

Ở khung tín hiệu có chiều dài 30 ms có kết quả như sau:

Đơn vị đo: Hz

Tín hiệu	Đo thủ công	F0 dùng hàm tự tương quan		F0 dùng hàm tự tương quan qua bộ lọc trung vị					
		F0	Độ lệch	N=3	Độ lệch	N=5	Độ lệch	N=7	Độ lệch
/a/	112,0	116,7	4,8	116,5	4,6	114,7	2,8	115,2	3,3
/e/	107,7	113,0	5,3	112,9	5,2	113,0	5,3	113,3	5,6
/i/	117,2	114,9	2,3	115,8	1,4	115,8	1,4	114,5	2,7
/o/	103,2	112,2	8,9	112,0	8,8	111,8	8,5	111,3	8,1
/u/	115,4	135,1	19,7	134,9	19,5	124,3	9,0	123,5	8,2

Bảng 3.5 - Khảo sát kích thước bộ lọc trung vị với một người nam

ở khung tín hiệu 30 ms

Đơn vị đo: Hz

Tín hiệu	Đo thủ công	F0 dùng hàm tự tương quan		F0 dùng hàm tự tương quan qua bộ lọc trung vị					
		F0	Độ lệch	N=3	Độ lệch	N=5	Độ lệch	N=7	Độ lệch
/a/	315,1	319,7	4,6	319,8	4,7	319,7	4,6	319,7	4,6
/e/	310,7	321,5	10,9	321,5	10,9	321,4	10,8	321,5	10,8
/i/	334,1	332,0	2,1	332,0	2,1	331,9	2,2	331,9	2,2
/o/	317,4	318,5	1,1	318,5	1,1	318,6	1,2	318,3	1,0
/u/	336,7	332,8	3,9	332,8	3,9	332,8	3,9	332,8	3,9

Bảng 3.6 - Khảo sát kích thước bộ lọc trung vị với một người nữ ở khung tín hiệu 30 ms

Qua bảng 3.11, đối với tín hiệu thu được ở người nữ, sự thay đổi kích thước của bộ lọc trung vị không cho kết quả chênh lệch nhiều và cũng không thể đánh giá được kích thước nào là tốt nhất.

Ở bảng 3.12, đối với tín hiệu thu được ở người nam cũng không thể đưa ra được kết luận kích thước của bộ lọc trung vị nào là tốt nhất.

Qua các bảng số liệu từ 3.7 đến bảng số liệu 3.12 có thể thấy rằng: ở khung tín hiệu 15 ms đối với giọng nữ, kích thước bộ lọc trung vị N = 3 là cho kết quả tốt nhất. Tuy nhiên, với độ dài khung là 15 ms, giọng nam thu được lại không thể tính được F0 trong hầu hết các trường hợp. Với khung tín hiệu có độ dài là 20 ms, kích thước bộ lọc trung vị N = 7 lại cho kết quả tốt nhất đối với giọng nam thu được. Với giọng nữ thu được thì không kết luận được kích thước của bộ lọc trung bị nào là tốt nhất. Khi khảo sát ở độ dài của khung tín hiệu là 30 ms, cũng không thể đưa ra được kết luận kích thước của bộ lọc trung bị nào là tốt nhất.

Vì tín hiệu đưa vào không thể biết được là giọng nam hay giọng nữ, nên các kết quả F0 qua bộ lọc trung vị, tôi sử dụng kích thước bộ lọc trung vị là N = 7.

3.6. Khảo sát ngưỡng xác định hữu thanh/vô thanh

Trong thuật toán tìm F0 dùng hàm tự tương quan, một trong những tham số quan trọng của thuật toán là ngưỡng biên độ cực đại cục bộ của hàm tự tương quan để xác định âm hữu thanh hay âm vô thanh. Việc tăng hoặc giảm ngưỡng này ảnh hưởng đến việc xác định âm hữu thanh hoặc âm vô thanh của đoạn tín hiệu tiếng nói (phần 2.4.2). Như đã trình bày trong chương 2, hàm tự tương quan $r(\tau)$ của một khung tín hiệu đạt

cực đại toàn cục tại giá trị độ trễ τ = 0. Trong phạm vi luận văn, tôi tiến hành khảo sát các ngưỡng có giá trị bằng 30%, 50% và 70% của giá trị cực đại toàn cục *r(0)* của hàm tự tương quan để tìm ra ngưỡng tối ưu.

Ngoài ra, một ràng buộc liên quan đến giá trị F0 cần tìm là F0 của tín hiệu tiếng nói của người trưởng thành thường dao động trong khoảng từ $F0_{min}$ = 80 Hz đến $F0_{max}$ = 400 Hz. Vì vậy, chỉ các điểm cực đại cục bộ của hàm tự tương quan có độ trễ τ trong khoảng từ $\tau_{min} = 1/F0_{max}$ = 0,0025 s đến $\tau_{max} = 1/F0_{min}$ = 0,0125 s mới được xét đến khi tìm chu kỳ cơ bản T0, sau đó suy ra F0, của tín hiệu.

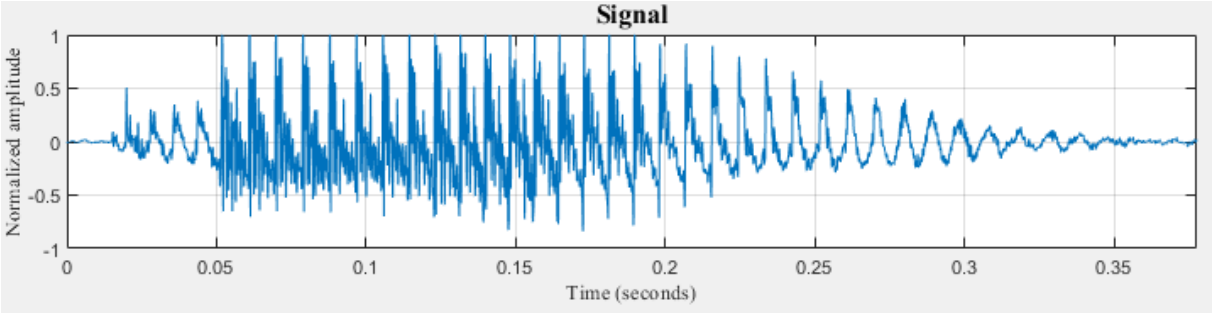
Để xác định được các điểm cực đại cục bộ (gọi là các đỉnh) của hàm tự tương quan của tín hiệu, tôi sử dụng hàm **findpeaks()** của Matlab với các tham số:

```
[y_peak,y_location]
=findpeaks(corr,'MinPeakDistance',mpd, 'MinPeakHeight',mph);
```

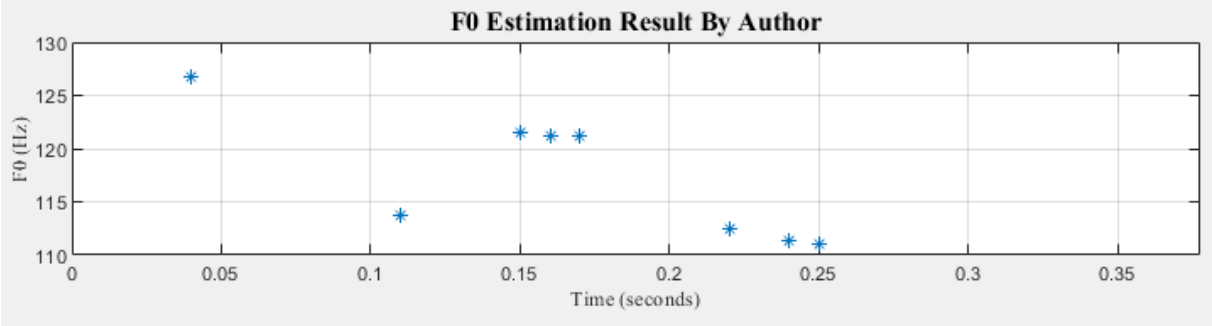
Trong đó, `corr` là hàm tự tương quan của tín hiệu ban đầu, `'MinPeakDistance'` là tuỳ chọn mô tả khoảng cách tối thiểu giữa các đỉnh, `'MinPeakHeight'` là tuỳ chọn mô tả biên độ tối thiểu của một đỉnh. Giá trị biên độ tối thiểu của một đỉnh (biến `mph`) chính là ngưỡng xác định âm hữu thanh/vô thanh. Còn giá trị khoảng cách tối thiểu giữa các đỉnh (biến `mpd`) chính là chu kỳ cơ bản nhỏ nhất mà thuật toán có thể xác định được: $T0_{min} = 1/F0_{max}$. Sau khi đã xác định được các thông số của hàm **findpeaks()**, tôi tiến hành tìm chu kỳ cơ bản T0 bằng cách xác định độ trễ τ^* ứng với đỉnh có biên độ lớn nhất trong số các đỉnh mà hàm **findpeaks()** tìm được. Nếu τ^* nằm trong khoảng từ τ_{min} đến τ_{max} thì giá trị τ^* chính là chu kỳ cơ bản T0 cần tìm, từ đó suy ra $F0 = 1/\tau^*$.

Để xác định giá trị ngưỡng tốt nhất, tôi khảo sát một cách định tính trên các tín hiệu của 3 giọng nam và 3 giọng nữ trong tập dữ liệu thử nghiệm. Hình 3.8 đến Hình 3.13 thể hiện các kết quả tính F0 đối với 3 giá trị ngưỡng bằng 30%, 50%, và 70% của *r(0)* cho mỗi giọng.

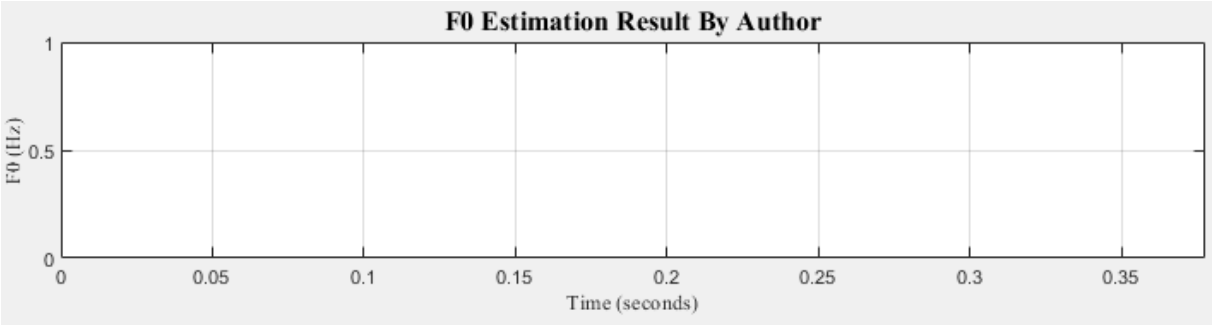
Với người nam thứ nhất, kết quả khảo sát như sau:



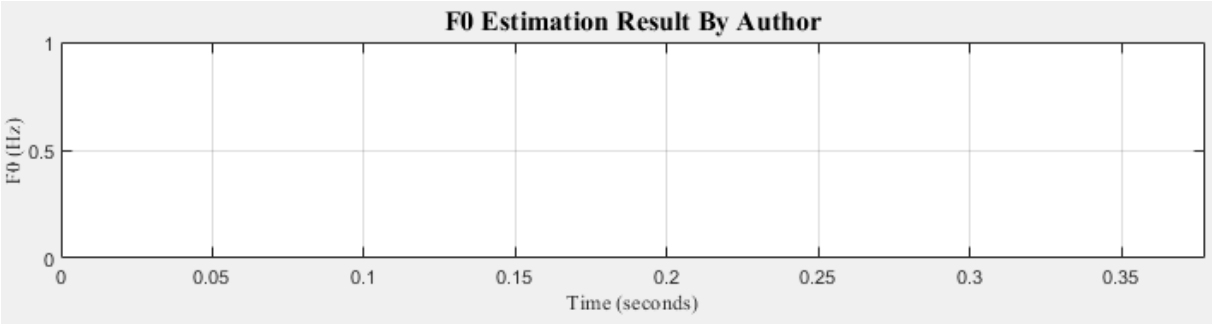
(a) Tín hiệu âm /a/ của người nam thứ nhất



(b) Đường F0 được tính với ngưỡng = 0.3*$r(0)$



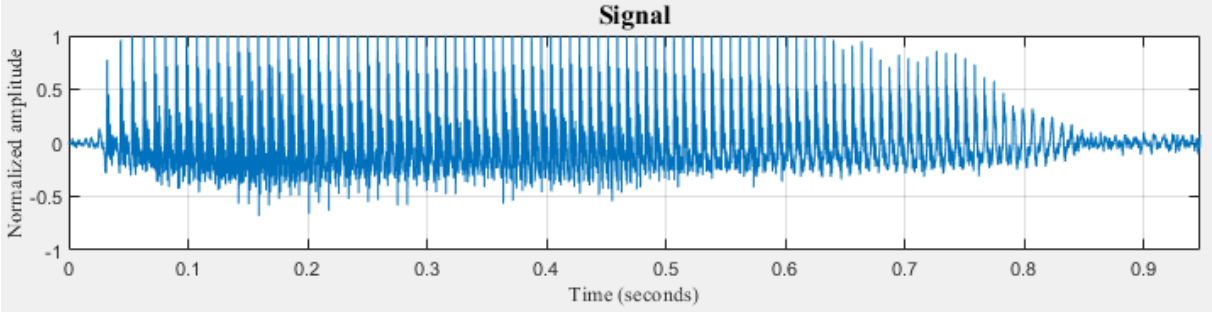
(c) Đường F0 được tính với ngưỡng = 0.5*$r(0)$



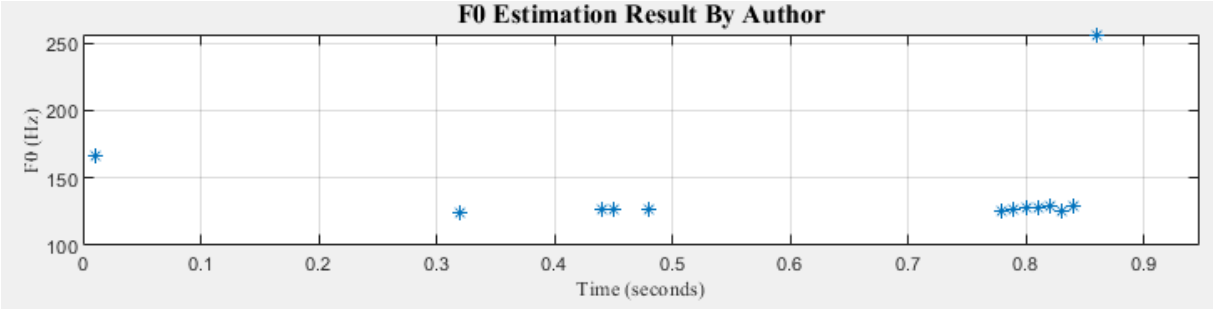
(d) Đường F0 được tính với ngưỡng = 0.7*$r(0)$

Hình 3.8 – Kết quả tính F0 của người nam thứ nhất theo các ngưỡng khác nhau

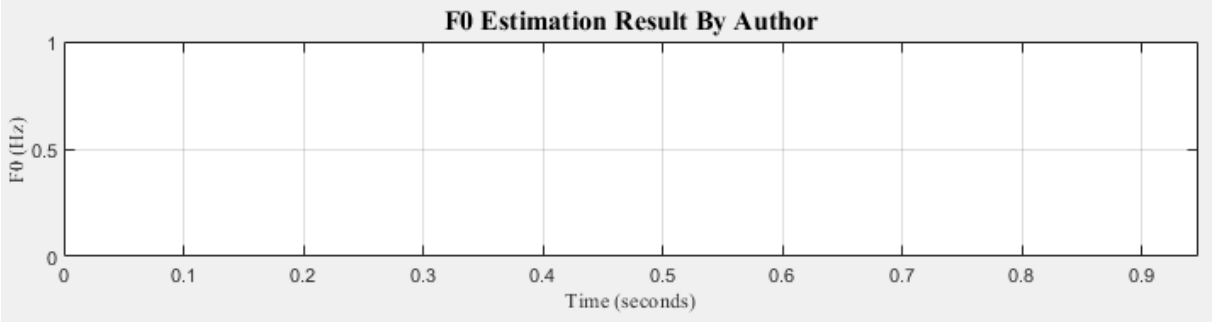
Với người nam thứ hai, kết quả khảo sát như sau:



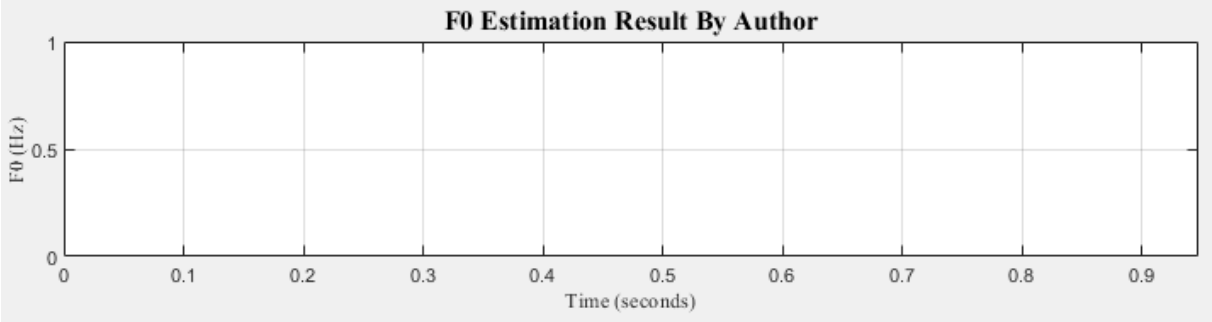
(a) Tín hiệu âm /a/ của người nam thứ hai



(b) Đường F0 được tính với ngưỡng = 0.3*r(0)



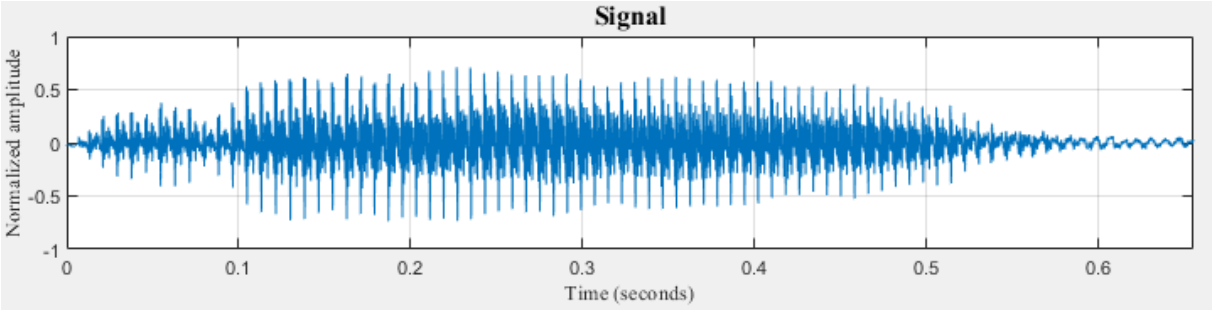
(c) Đường F0 được tính với ngưỡng = 0.5*r(0)



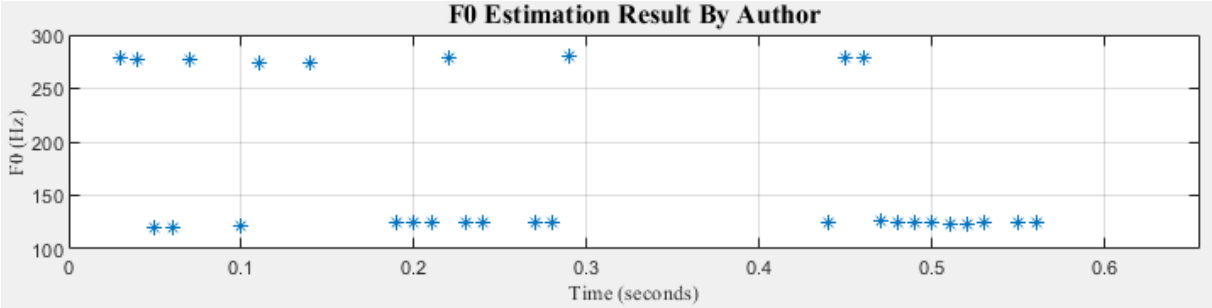
(d) Đường F0 được tính với ngưỡng = 0.7*r(0)

Hình 3.9 – Kết quả tính F0 của người nam thứ hai theo các ngưỡng khác nhau

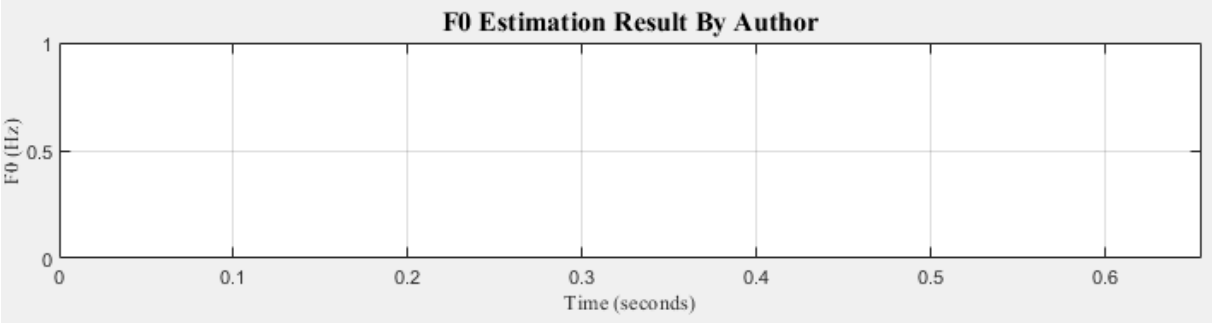
Với người nam thứ ba, kết quả khảo sát như sau:



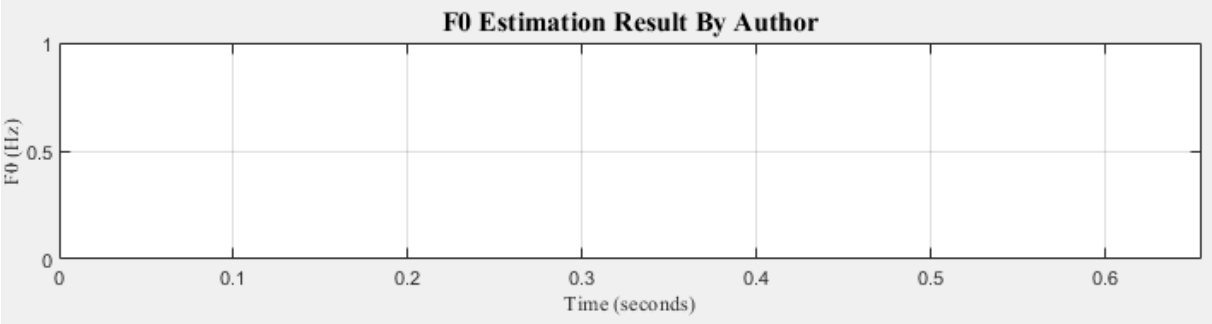
(a) Tín hiệu âm /a/ của người nam thứ ba



(b) Đường F0 được tính với ngưỡng = 0.3*r(0)



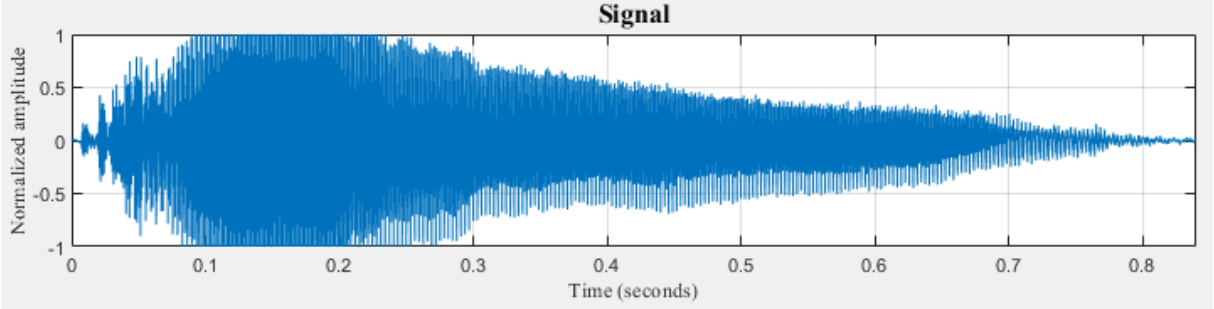
(c) Đường F0 được tính với ngưỡng = 0.5*r(0)



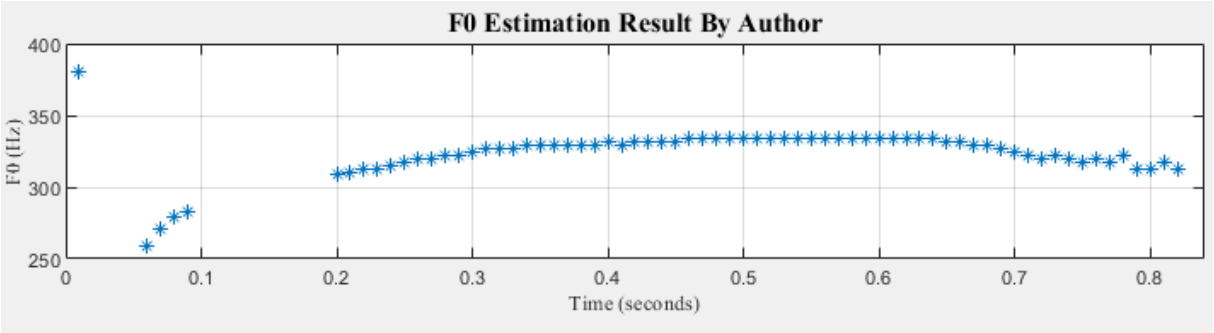
(d) Đường F0 được tính với ngưỡng = 0.7*r(0)

Hình 3.10 – Kết quả tính F0 của người nam thứ ba theo các ngưỡng khác nhau

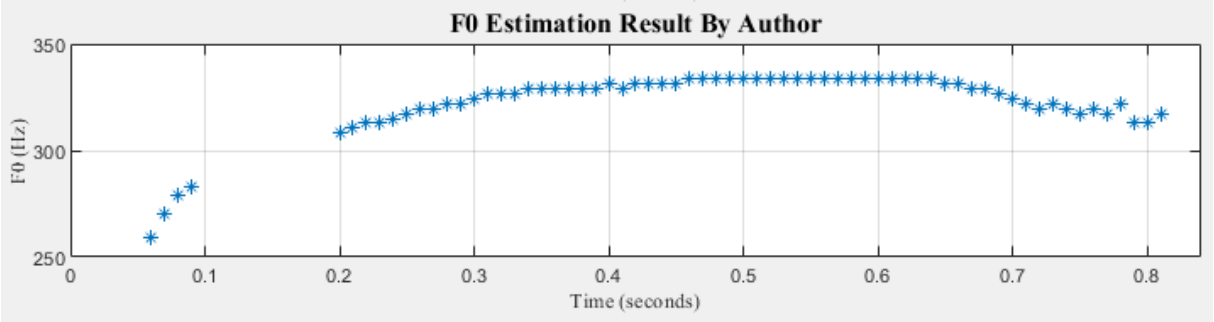
Với người nữ thứ nhất, kết quả khảo sát như sau:



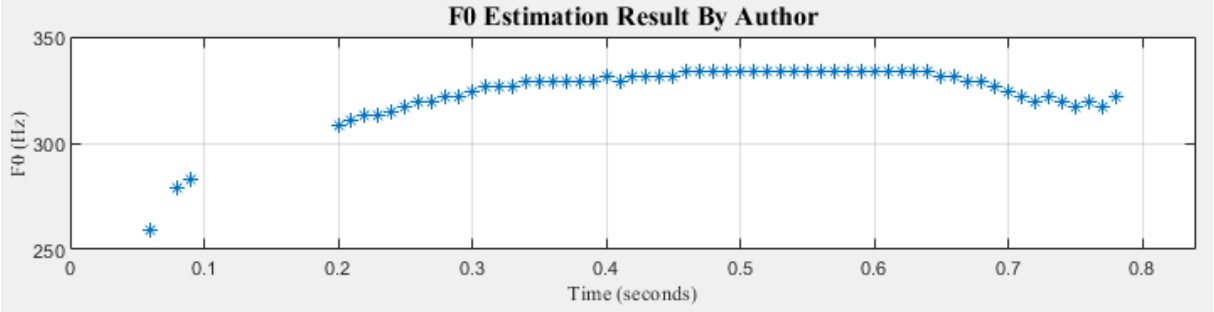
(a) Tín hiệu âm /a/ của người nữ thứ nhất



(b) Đường F0 được tính với ngưỡng = 0.3*$r(0)$



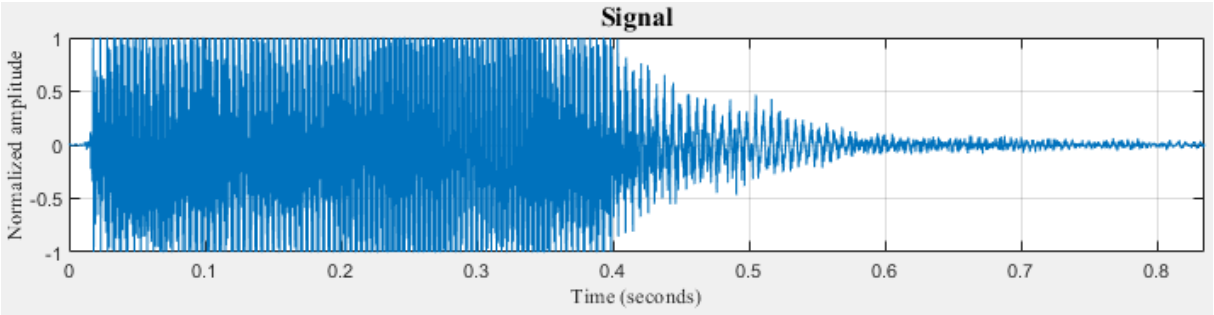
(c) Đường F0 được tính với ngưỡng = 0.5*$r(0)$



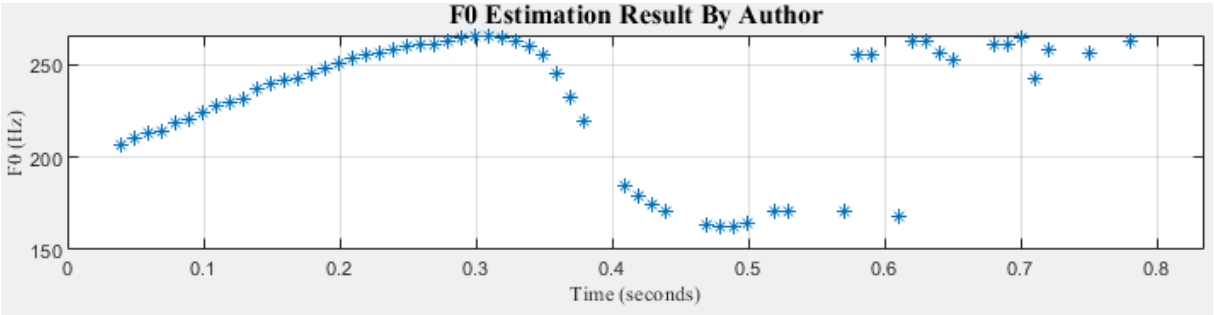
(d) Đường F0 được tính với ngưỡng = 0.7*$r(0)$

Hình 3.11 – Kết quả tính F0 của người nữ thứ nhất theo các ngưỡng khác nhau

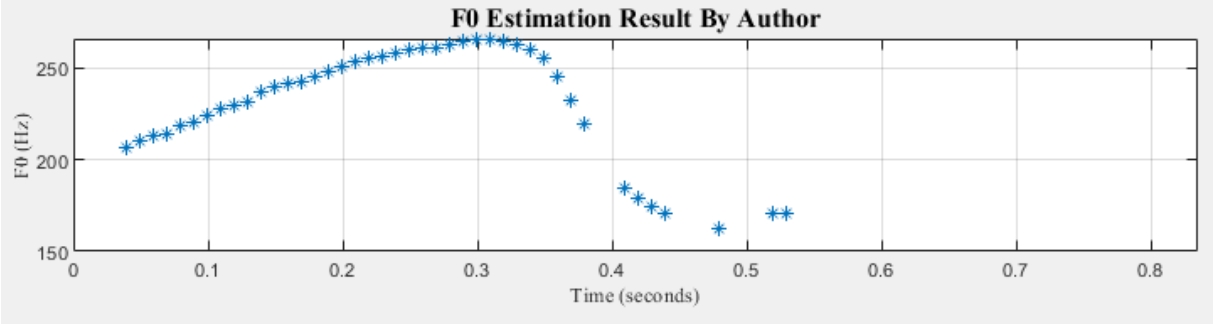
Với người nữ thứ hai, kết quả khảo sát như sau:



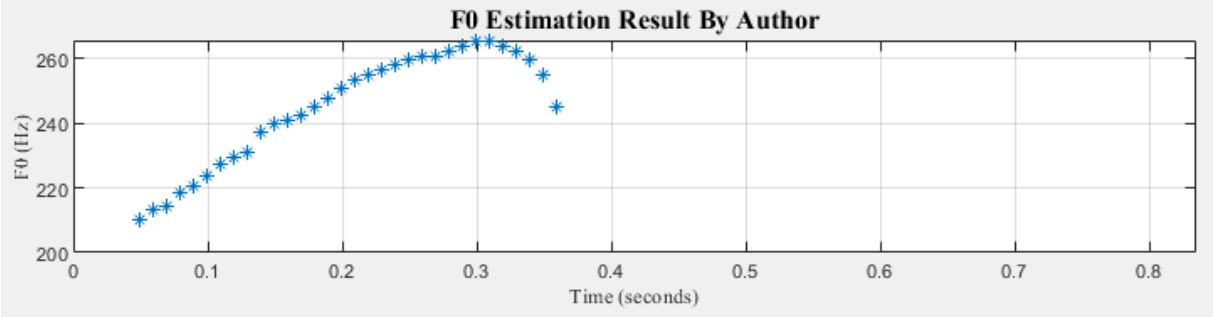
(a) Tín hiệu âm /a/ của người nữ thứ hai



(b) Đường F0 được tính với ngưỡng = 0.3*$r(0)$



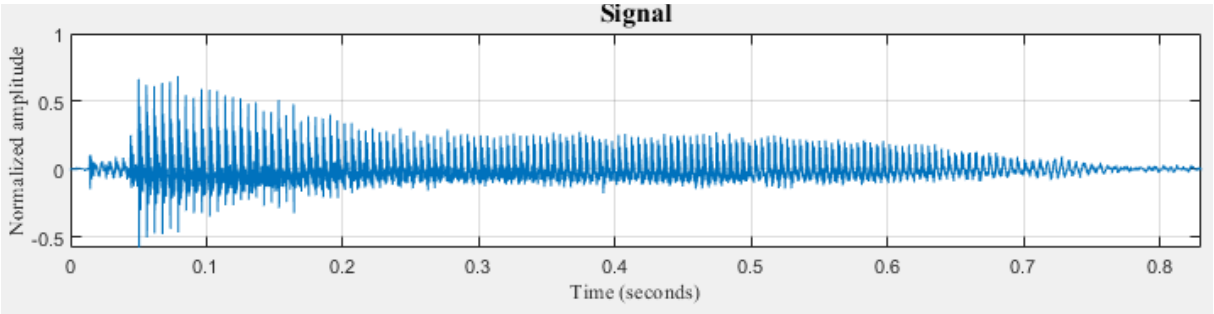
(c) Đường F0 được tính với ngưỡng = 0.5*$r(0)$



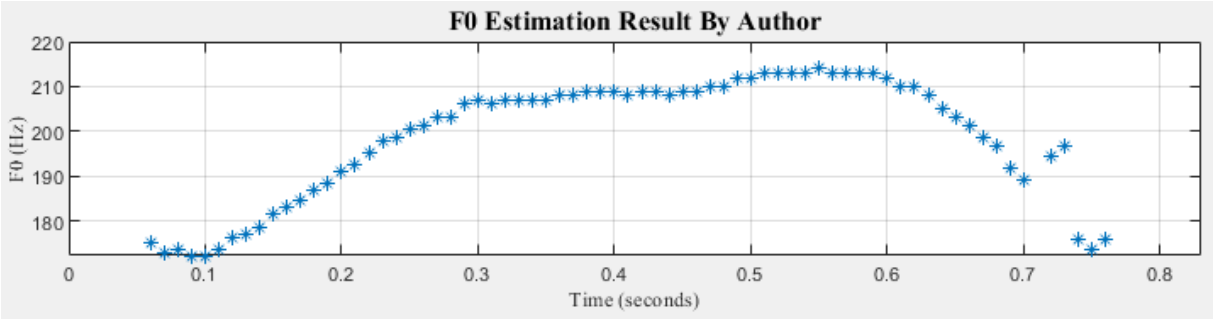
(d) Đường F0 được tính với ngưỡng = 0.7*$r(0)$

Hình 3.12 – Kết quả tính F0 của người nữ thứ hai theo các ngưỡng khác nhau

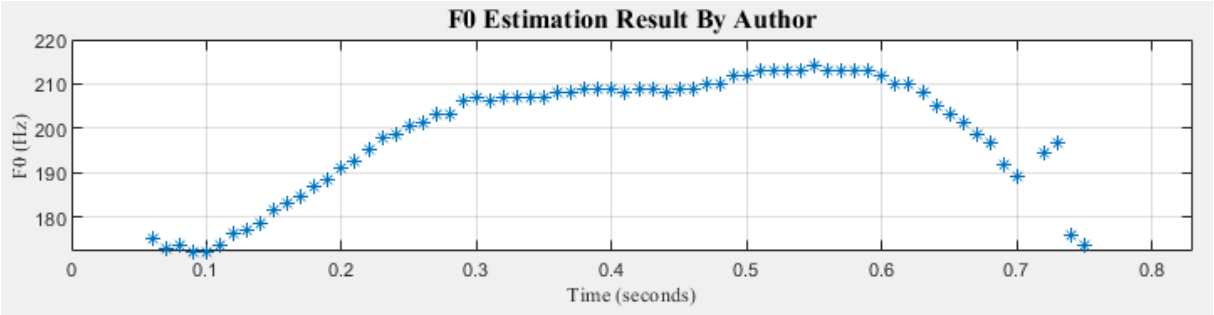
Với người nữ thứ ba, kết quả khảo sát như sau:



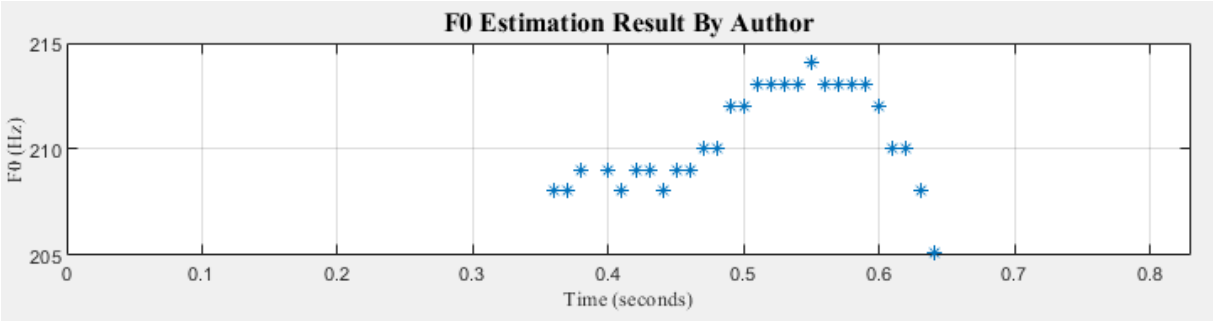
(a) Tín hiệu âm /a/ của người nữ thứ ba



(b) Đường F0 được tính với ngưỡng = 0.3*$r(0)$



(c) Đường F0 được tính với ngưỡng = 0.5*$r(0)$



(d) Đường F0 được tính với ngưỡng = 0.7*$r(0)$

Hình 3.13 – Kết quả tính F0 của người nữ thứ ba theo các ngưỡng khác nhau

Qua các kết quả khảo sát định tính trên đối với 3 ngưỡng xác định hữu thanh/vô thanh, có thể nhận xét là ngưỡng 30% của $r(0)$ có lỗi xác định hữu thanh/vô thanh thấp nhất ở cả giọng nam và nữ so với 2 giá trị ngưỡng còn lại. Điều này thể hiện ở chỗ các kết quả tính F0 với ngưỡng 30% của $r(0)$ luôn cho ra nhiều giá trị F0 xác định nhất, nghĩa là có nhiều khung tín hiệu được cho là tuần hoàn nhất, khớp với thực tế là các tín hiệu được khảo sát đều là nguyên âm. Ngược lại, ngưỡng 70% của $r(0)$ luôn cho ra ít giá trị F0 xác định nhất, nghĩa là có nhiều khung tín hiệu tuần hoàn bị thuật toán xác định nhầm là không tuần hoàn. Các kết quả này khớp với thảo luận trong phần 2.4.2.

Với kết quả khảo sát được, tôi chọn ngưỡng bằng 30% của $r(0)$ để xác định F0 của tín hiệu tiếng nói trong phần còn lại của luận văn.

3.7. So sánh cài đặt hàm tự tương quan tự làm với hàm của Matlab

Việc tính toán kết quả đo tần số cơ bản của hàm tự tương quan của tác giả với hàm tự tương quan (TTQ) của Matlab được thực hiện trên ba độ dài khung tín hiệu: 15 ms, 20 ms và 30 ms. Vì F0 đo được của tín hiệu thu được là một dãy các giá trị trên các khung hữu thanh, nên khi tiến hành khảo sát độ lệch giữa hàm tự tương quan của tác giả và hàm tự tương quan của Matlab, tôi sử dụng giá trị trung bình để xác định F0 cuối cùng. Đây cũng là cách tôi sử dụng cho các kết quả khác trong luận văn khi tính F0 của tín hiệu tiếng nói.

Đối với giọng của một người nam, với độ dài khung là 15 ms, kết quả đo được sau khi lọc trung vị như sau:

Đơn vị đo: Hz

Tín hiệu	F0 tính bằng hàm TTQ của tác giả	F0 tính bằng hàm TTQ của Matlab	Độ lệch của hai giá trị F0
/a/	296,0	294,0	2,0
/e/	230,9	229,7	1,2
/i/	187,3	186,4	0,9
/o/	N/A	N/A	N/A
/u/	336,2	333,6	2,6

Bảng 3.7 - Kết quả tính F0 (Hz) với độ dài khung 15 ms của một người nam

Với độ dài khung tín hiệu xử lý ngắn hạn là 20 ms, kết quả đo được như sau:

Đơn vị đo: Hz

Tín hiệu	F0 tính bằng hàm TTQ của tác giả	F0 tính bằng hàm TTQ của Matlab	Độ lệch của hai giá trị F0
/a/	112,8	112,5	0,3
/e/	114,0	113,7	0,3
/i/	117,8	117,5	0,3
/o/	114,0	113,7	0,3
/u/	124,2	123,8	0,4

Bảng 3.8 - Kết quả tính F0 (Hz) với độ dài khung 20 ms của một người nam

Và ở độ dài khung tín hiệu xử lý ngắn hạn là 30 ms, kết quả đo được như sau:

Đơn vị đo: Hz

Tín hiệu	F0 tính bằng hàm TTQ của tác giả	F0 tính bằng hàm TTQ của Matlab	Độ lệch của hai giá trị F0
/a/	114,8	114,5	0,3
/e/	112,7	112,5	0,2
/i/	115,1	114,8	0,3
/o/	111,3	111,1	0,2
/u/	121,5	121,1	0,4

Bảng 3.9 - Kết quả tính F0 (Hz) với độ dài khung 30 ms của một người nam

Các Bảng 3.1, 3.2 và 3.3 cho thấy, với các độ dài khung khác nhau, đối với giọng nói của người nam, kết quả đo được giữa hàm tự tương quan của tác giả và hàm tự tương quan của Matlab không chênh lệch nhiều (độ lệch đều nhỏ hơn 3 Hz).

Để kiểm chứng thuật toán trên giọng nữ, tôi tiến hành tương tư như với giọng nam và được các kết quả như sau:

Đơn vị đo: Hz

Tín hiệu	F0 tính bằng hàm TTQ của tác giả	F0 tính bằng hàm TTQ của Matlab	Độ lệch của hai giá trị F0
/a/	325,0	322,6	2,4
/e/	322,2	319,9	2,3
/i/	333,6	330,9	2,7
/o/	320,7	318,4	2,3
/u/	331,9	329,4	2,5

Bảng 3.10 - Kết quả tính F0 (Hz) với độ dài khung 15 ms của một người nữ

Đơn vị đo: Hz

Tín hiệu	F0 tính bằng hàm TTQ của tác giả	F0 tính bằng hàm TTQ của Matlab	Độ lệch của hai giá trị F0
/a/	323,1	320,7	2,4
/e/	321,3	319,0	2,3
/i/	333,3	330,8	2,5
/o/	320,5	318,1	2,4
/u/	332,8	330,3	2,5

Bảng 3.11 - Kết quả tính F0 (Hz) với độ dài khung 20 ms của một người nữ

Đơn vị đo: Hz

Tín hiệu	F0 tính bằng hàm TTQ của tác giả	F0 tính bằng hàm TTQ của Matlab	Độ lệch của hai giá trị F0
/a/	319,3	317,4	1,9
/e/	318,1	315,8	2,3
/i/	332,0	329,5	2,5
/o/	318,5	316,2	2,3
/u/	332,6	330,1	2,5

Bảng 3.12 - Kết quả tính F0 (Hz) với độ dài khung 30 ms của một người nữ

Đối với giọng của một người nữ, hàm tự tương quan do tác giả viết và hàm tự tương quan của Matlab cũng không có sự chênh lệch nhiều về cách tính F0 (độ lệch đều nhỏ hơn 3 Hz).

Qua các bảng số liệu so sánh giữa hàm tự tương quan của tác giả và hàm tự tương quan của Matlab cho thấy có thể sử dụng hàm tự tương quan của tác giả để tính toán giá

trị F0 đối với tín hiệu đưa vào khảo sát. Trong các số liệu trở về sau của luận văn, tôi sử dụng hàm tự tương quan tự viết để đánh giá thuật toán tính F0.

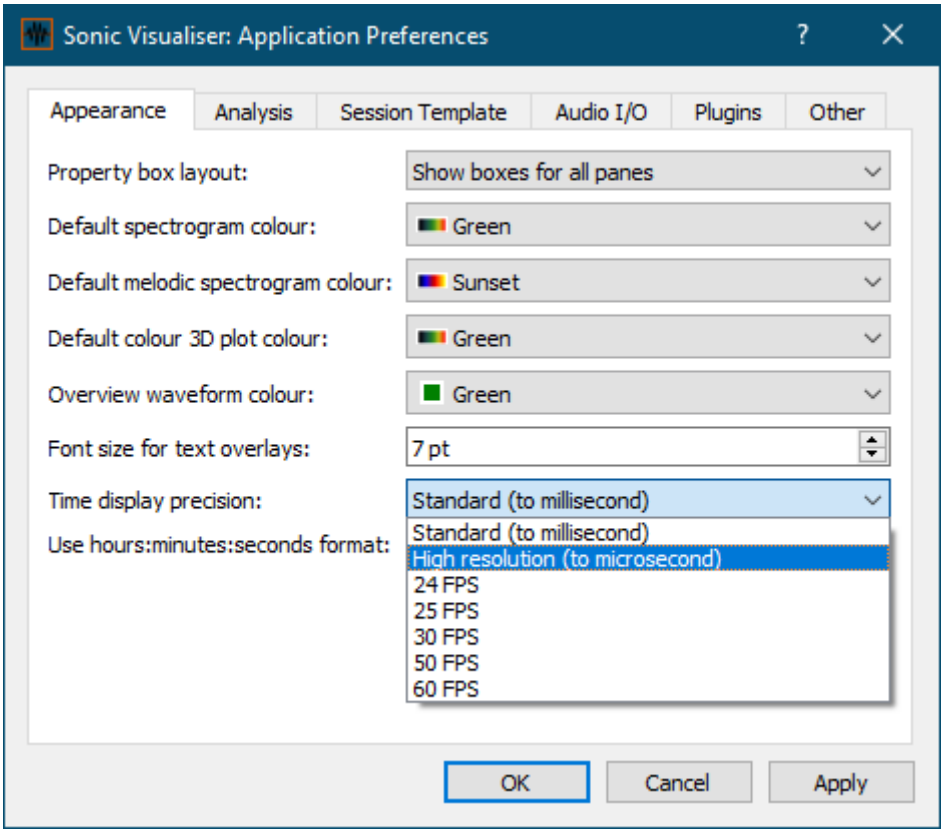
3.8. So sánh thuật toán tính F0 tự động với cách đo F0 thủ công

Trong phần này, tôi đánh giá sai số của thuật toán tính F0 tự cài đặt bằng cách tính độ lệch tuyệt đối giữa giá trị F0 chuẩn được đo thủ công và giá trị F0 tự động tính bởi thuật toán. Việc đánh giá sai số được thực hiện với các độ dài khung tín hiệu khác nhau nhằm tìm ra giá trị phù hợp nhất của tham số này.

3.8.1. Cách đo F0 thủ công

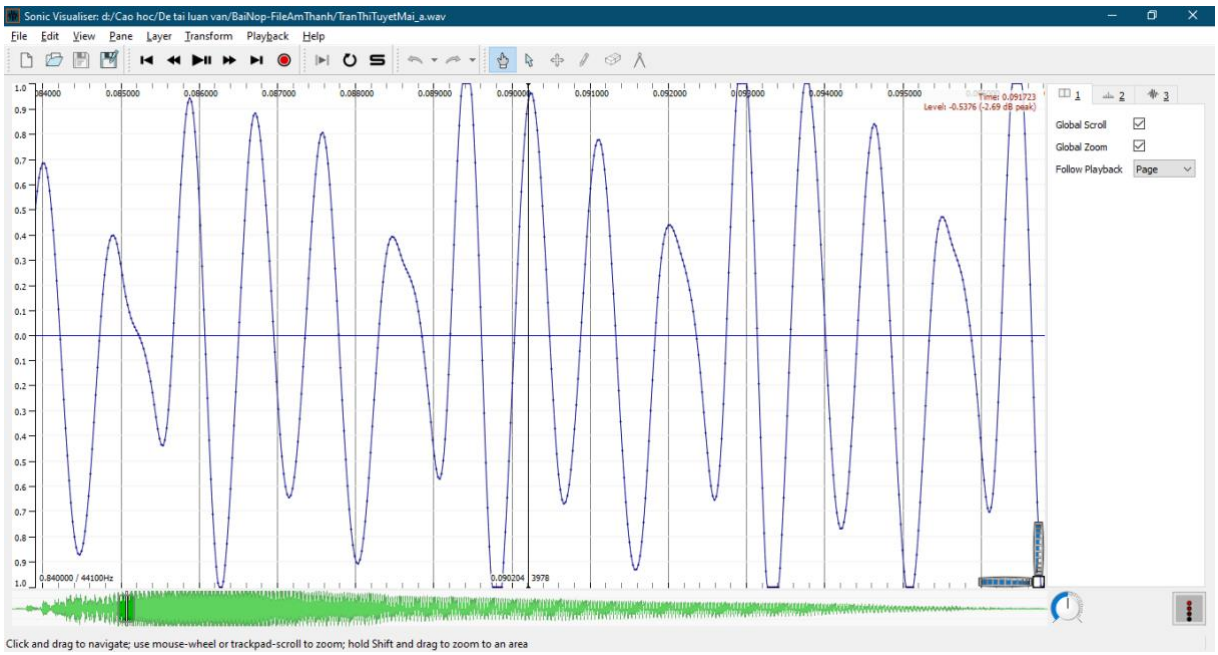
Để đo được F0 thủ công, tôi sử dụng công thức (2.4) và phần mềm Sonic Visualiser phiên bản 3.3. Đây là phần mềm miễn phí và có giấy phép GNU General Public License (phiên bản 2 hoặc lớn hơn).

Để đo được F0 của tín hiệu chính xác đến 1/10 Hz, cần phải chỉnh lại độ chính xác của khoảng cách cần lấy của phần mềm từ Standard (to milisecond) về High resolution (to microsecond) trong hộp hội thoại Preferences ở menu File (Hình 3.14).



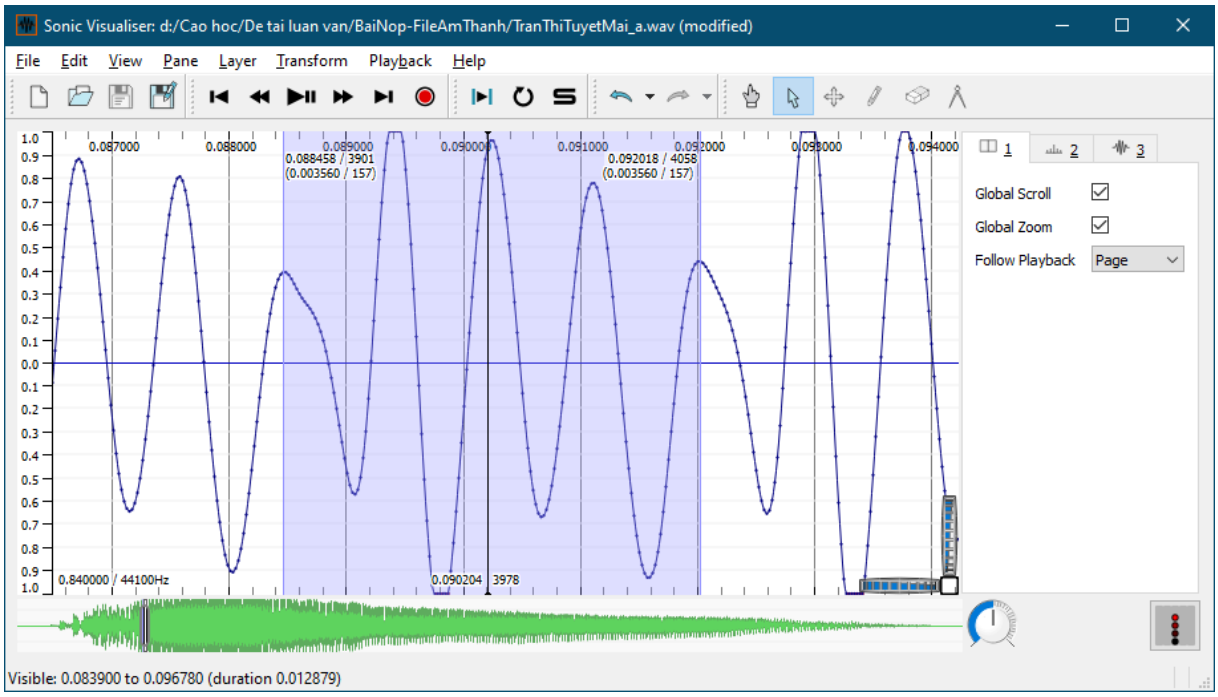
Hình 3.14 – Chuyển đổi độ chính xác khi đo trong phần mềm Sonic Visualiser

Trước tiên, cần phải chọn file tín hiệu âm thanh cần đo. Sau đó thực hiện phóng to đoạn tín hiệu cần đo để nhìn thấy rõ vài chu kỳ của tín hiệu (Hình 3.15).



Hình 3.15 – Phóng to đoạn tín hiệu trong phần mềm Sonic Visualiser

Dùng công cụ Select trong phần mềm để thực hiện đo chu kỳ cơ bản của tín hiệu (Hình 3.16).



Hình 3.16 – Đo chu kỳ cơ bản của tín hiệu bằng phần mềm Sonic Visualiser

Trên hình 3.20, chu kỳ cơ bản của tín hiệu đo được là T0 = 0,003560, từ đó suy ra F0 = 1/T0 ≈ 280,9 Hz.

Trong phạm vi của luận văn, tôi thực hiện đo ngẫu nhiên một đoạn tín hiệu 5 lần để xác định T0. Từ các giá trị T0 thu được, tôi tính giá trị trung bình để lấy kết quả F0 cuối cùng làm giá trị chuẩn.

3.8.2. Kết quả đối với giọng nam

Tôi tiến hành thực hiện trên ba giọng nam thu được. Và tiến hành lần lượt khảo sát giá trị tính F0 của hàm tự tương quan kết hợp với lọc trung vị qua độ dài khung là 15 ms, 20 ms, và 30 ms.

Với độ dài khung 15 ms, kết quả tính được như sau:

Đơn vị đo: Hz

Tín hiệu	F0 đo thủ công	F0 tính tự động	Độ lệch của hai giá trị F0
/a/	112,0	296,0	184,0
/e/	107,7	230,9	123,2
/i/	117,2	187,3	93,5
/o/	103,2	N/A	N/A
/u/	115,4	336,2	220,8
Trung bình	**111,1**	**262,6**	**149,5**

Bảng 3.13 – Kết quả đo F0 với độ dài khung 15 ms của người nam thứ nhất

Đơn vị đo: Hz

Tín hiệu	F0 đo thủ công	F0 tính tự động	Độ lệch của hai giá trị F0
/a/	119,9	N/A	N/A
/e/	126,2	N/A	N/A
/i/	127,9	205,1	77,3
/o/	132,2	207,7	75,5
/u/	135,4	138,5	3,0
Trung bình	**128,3**	**183,8**	**51,9**

Bảng 3.14 – Kết quả đo F0 với độ dài khung 15 ms của người nam thứ hai

Đơn vị đo: Hz

Tín hiệu	F0 đo thủ công	F0 tính tự động	Độ lệch của hai giá trị F0
/a/	122,2	262,1	139,9
/e/	122,2	N/A	N/A
/i/	126,2	378,9	252,8
/o/	122,0	366,8	244,8
/u/	126,0	378,1	252,0
Trung bình	**123,7**	**346,5**	**222,4**

Bảng 3.15 – Kết quả đo F0 với độ dài khung 15 ms của người nam thứ ba

Bảng 3.13, 3.14 và 3.15 cho thấy, nếu khung tín hiệu có độ dài quá ngắn, việc tính F0 bằng hàm tự tương quan sẽ có sai số rất lớn hoặc có thể sẽ không tính toán được F0 (như ở âm /o/ ở bảng 3.13, âm /i/ và âm /o/ ở bảng 3.14, các âm ở bảng 3.15). Nguyên nhân là do chu kỳ cơ bản của giọng nam thường dài từ 10 ms (đối với F0 = 100 Hz) đến 5 ms (đối với F0 = 200 Hz). Nếu khung tín hiệu dài 15 ms thì chỉ chứa khoảng 1,5 chu kỳ của tín hiệu đối với người nói có F0 xấp xỉ 100 Hz trong các trường hợp trên. Điều

này làm cho thuật toán tính sai hoặc không tính được giá trị F0 do cần ít nhất 2 chu kỳ liên tiếp của tín hiệu để hàm tự tương quan thể hiện được tính tuần hoàn của khung tín hiệu.

Với khung tín hiệu ngắn hạn có độ dài 20 ms, kết quả tính được như sau:

Đơn vị đo: Hz

Tín hiệu	F0 đo thủ công	F0 tính tự động	Độ lệch của hai giá trị F0
/a/	112,0	115,2	3,3
/e/	107,7	114,5	6,8
/i/	117,2	118,6	1,3
/o/	103,2	113,1	9,8
/u/	115,4	123,0	7,6
Trung bình	**111,1**	**116,9**	**5,8**

Bảng 3.16 – Kết quả đo F0 với độ dài khung 20 ms của người nam thứ nhất

Đơn vị đo: Hz

Tín hiệu	F0 đo thủ công	F0 tính tự động	Độ lệch của hai giá trị F0
/a/	119,9	126,5	6,6
/e/	126,2	129,5	3,3
/i/	127,9	156,9	29,0
/o/	132,2	143,1	10,9
/u/	135,4	136,5	1,1
Trung bình	**128,3**	**138,5**	**10,2**

Bảng 3.17 – Kết quả đo F0 với độ dài khung 20 ms của người nam thứ hai

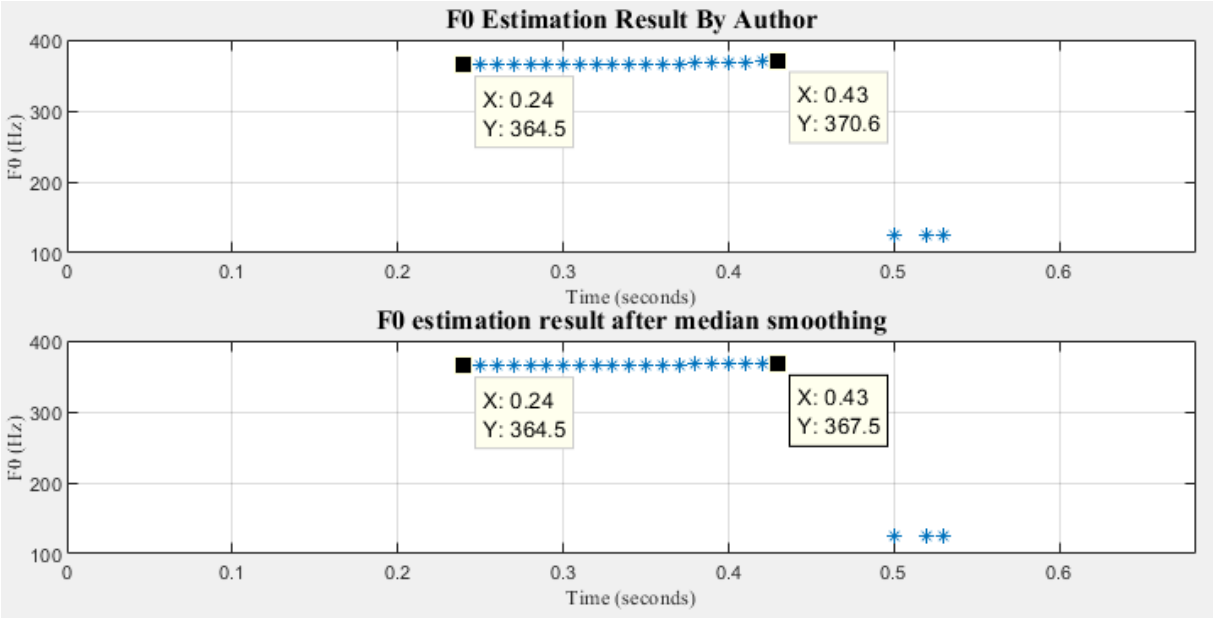
Đơn vị đo: Hz

Tín hiệu	F0 đo thủ công	F0 tính tự động	Độ lệch của hai giá trị F0
/a/	122,2	134,6	12,5
/e/	122,2	158,4	36,1
/i/	126,2	325,9	199,7
/o/	122,0	334,1	212,1
/u/	126,0	337,8	211,8
Trung bình	**123,7**	**239,2**	**115,0**

Bảng 3.18 – Kết quả đo F0 với độ dài khung 20 ms của người nam thứ ba

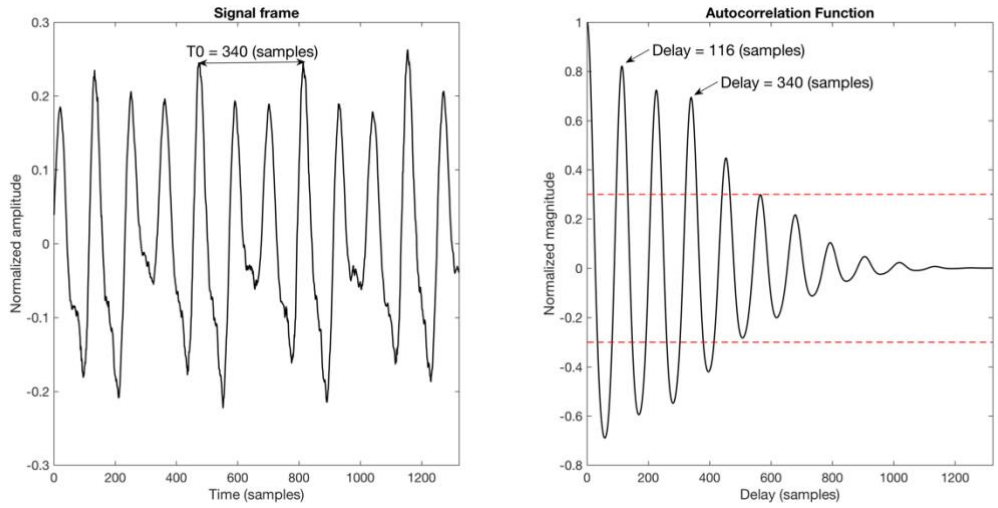
Qua bảng 3.16 cho thấy độ lệch giá trị F0 tính được của người nam thứ nhất đo được dao động dưới 10 Hz. Ở bảng 3.17, kết quả tính F0 thủ công so với cách tính F0 tự động có sự chênh lệch nhiều hơn. Tuy nhiên sự chênh lệch này là không đáng kể.

Bảng 3.18 cho thấy độ lệch giữa cách đo F0 thủ công và F0 tự động tăng đột biến. Đặc biệt với các âm /i/, /o/, và /u/, giá trị F0 tính bằng thuật toán cao gấp gần 3 lần giá trị F0 chuẩn đo thủ công (gọi là lỗi cao độ ảo [2]). Điều này thường do tín hiệu có hình dạng tuần hoàn một cách bất thường.

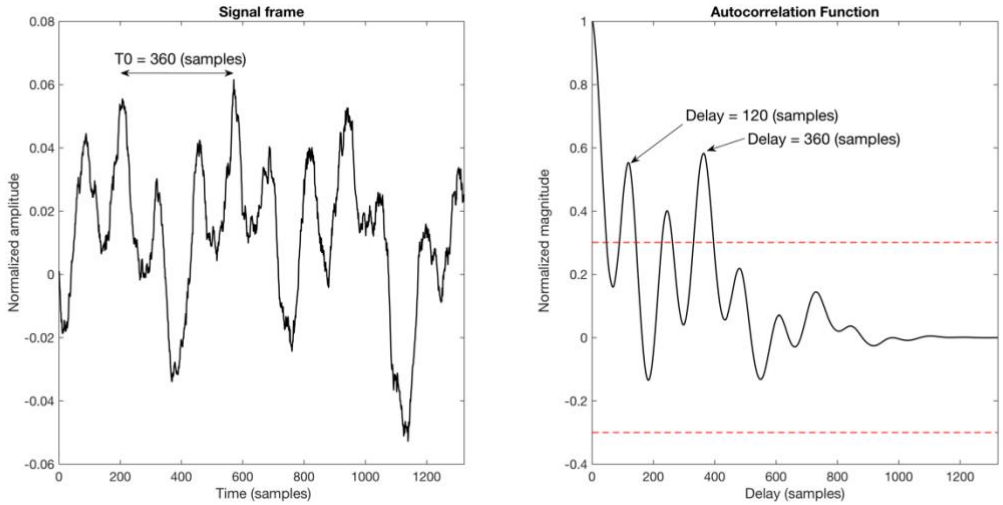


Hình 3.17 – Kết quả đo F0 của tín hiệu âm /o/ với độ dài khung 20 ms của người nam thứ ba

Xét kết quả tính F0 của tín hiệu âm /u/ của người nam thứ ba trong Hình 3.17. Kết quả tính F0 tự động cho thấy các giá trị F0 của các khung tín hiệu ở trước thời điểm 0,45 giây đều xấp xỉ 360 Hz. Trong khi đó, các khung tín hiệu từ thời điểm 0,5 giây trở về sau có F0 xấp xỉ 120 Hz (giá trị chuẩn đối với giọng nam này). Quan sát một khung tín hiệu ở trước thời điểm 0,45 giây và hàm tự tương quan của khung này (Hình 3.18), đồ thị tín hiệu cho thấy một chu kỳ cơ bản thực sự của tín hiệu (T0 = 340 mẫu) trông giống như gồm 3 chu kỳ ảo, dẫn đến hàm tự tương quan đạt đỉnh cao nhất ở độ trễ xấp xỉ 1/3 của T0 (=116 mẫu) thay vì ở độ trễ bằng chính T0. Ngược lại, một khung tín hiệu ở sau thời điểm 0,5 giây và hàm tự tương quan của khung này (Hình 3.19) cho thấy, chu kỳ cơ bản thực sự của tín hiệu (T0 = 360 mẫu) thể hiện rõ hơn nhiều, dẫn đến hàm tự tương quan đạt đỉnh cao nhất ở đúng độ trễ bằng T0.



Hình 3.18 – Một khung tín hiệu bị lỗi cao độ ảo và hàm tự tương quan của nó



Hình 3.19 – Một khung tín hiệu không bị lỗi cao độ ảo và hàm tự tương quan của nó

Với khung tín hiệu có độ dài 30ms, kết quả tính được như sau:

Đơn vị đo: Hz

Tín hiệu	**F0 đo thủ công**	**F0 tính tự động**	**Độ lệch của hai giá trị F0**
/a/	112,0	115,2	3,3
/e/	107,7	113,3	5,6
/i/	117,2	114,5	2,7
/o/	103,2	111,3	8,1
/u/	115,4	123,0	7,6
Trung bình	**111,1**	**115,5**	**5,5**

Bảng 3.19 – Kết quả đo F0 với độ dài khung 30 ms của người nam thứ nhất

Đơn vị đo: Hz

Tín hiệu	**F0 đo thủ công**	**F0 tính tự động**	**Độ lệch của hai giá trị F0**
/a/	119,9	123,3	3,4
/e/	126,2	127,0	0,8

/i/	127,9	133,5	5,6
/o/	132,2	135,5	3,3
/u/	135,4	134,9	0,5
Trung bình	**128,3**	**130,8**	**2,7**

Bảng 3.20 – Kết quả đo F0 với độ dài khung 30 ms của người nam thứ hai

Đơn vị đo: Hz

Tín hiệu	**F0 đo thủ công**	**F0 tính tự động**	**Độ lệch của hai giá trị F0**
/a/	122,2	123,7	1,5
/e/	122,2	124,3	2,1
/i/	126,2	276,2	150,1
/o/	122,0	122,8	0,8
/u/	126,0	300,0	174,0
Trung bình	**123,7**	**206,0**	**82,3**

Bảng 3.21 – Kết quả đo F0 với độ dài khung 30 ms của người nam thứ ba

Các kết quả ở bảng 3.19 và bảng 3.20 cho thấy độ lệch giữa các tính F0 thủ công và cách tính F0 tự động bằng hàm tự tương quan thấp hơn so với các kết quả độ lệch tính được ở độ dài khung 20 ms. Tuy nhiên, đối với người nam thứ ba (bảng 3.21), lỗi cao độ ảo vẫn xuất hiện với âm /i/ và /u/ do dao động rung của dây thanh diễn ra khá bất thường trong quá trình phát âm của người này.

Kết luận: từ các kết quả đo F0 tự động trên các tín hiệu của 3 người nam đã khảo sát, độ chính xác của kết quả tính F0 tăng dần khi độ dài khung tín hiệu tăng từ 15 ms đến 30 ms. Thuật toán tự tương quan cũng cho thấy nó khá dễ mắc lỗi cao độ ảo khi người nói phát ra tín hiệu có tính chất tuần hoàn bất thường (ví dụ như người nam thứ ba).

3.8.3. Kết quả đối với giọng nữ

Tương tự như đối với giọng nam, việc đo F0 trên các giọng nữ cũng thực hiện lần lượt ở các độ dài khung 15 ms, 20 ms, 30 ms.

Với độ dài khung là 15 ms, tôi thu được kết quả như sau:

Đơn vị đo: Hz

Tín hiệu	**F0 đo thủ công**	**F0 tính tự động**	**Độ lệch của hai giá trị F0**
/a/	315,1	325,0	9,9
/e/	310,7	322,2	11,5
/i/	334,1	333,6	0,5
/o/	317,4	320,7	3,3
/u/	336,7	331,9	4,8
Trung bình	**322,8**	**326,7**	**6,0**

Bảng 3.22 – Kết quả đo F0 với độ dài khung 15 ms của người nữ thứ nhất

Đơn vị đo: Hz

Tín hiệu	F0 đo thủ công	F0 tính tự động	Độ lệch của hai giá trị F0
/a/	199,6	237,6	38,0
/e/	184,5	230,1	45,5
/i/	226,2	254,4	28,2
/o/	179,3	229,7	50,4
/u/	239,7	265,7	26,0
Trung bình	**205,9**	**243,5**	**37,6**

Bảng 3.23 – Kết quả đo F0 với độ dài khung 15 ms của người nữ thứ hai

Đơn vị đo: Hz

Tín hiệu	F0 đo thủ công	F0 tính tự động	Độ lệch của hai giá trị F0
/a/	188,6	198,5	9,9
/e/	203,9	216,0	12,1
/i/	215,0	217,6	2,6
/o/	195,8	209,7	13,9
/u/	213,0	242,6	29,6
Trung bình	**203,3**	**216,9**	**13,6**

Bảng 3.24 – Kết quả đo F0 với độ dài khung 15 ms của người nữ thứ ba

Có thể nhận thấy việc tính F0 dùng độ dài khung 15 ms đối với tín hiệu tiếng nói của giọng nữ (3 bảng 3.22, 3.23 và 3.24) có sai số ít hơn nhiều so với giọng nam (3 bảng 3.13, 3.14 và 3.15). Nguyên nhân là do chu kỳ cơ bản của giọng nữ thường dài từ 5 ms (đối với F0 = 200 Hz) đến 3 ms (đối với F0 = 333,3 Hz) nên với độ dài khung 15 ms (chứa từ 3 đến 5 chu kỳ liên tiếp của tín hiệu), thuật toán tự tương quan có khả năng xác định chu kỳ cơ bản T0 và suy ra F0 khá dễ dàng.

Với độ dài khung là 20 ms, kết quả thu được như sau:

Đơn vị đo: Hz

Tín hiệu	F0 đo thủ công	F0 tính tự động	Độ lệch của hai giá trị F0
/a/	315,1	323,1	8,0
/e/	310,7	321,3	10,7
/i/	334,1	333,3	0,8
/o/	317,4	320,5	3,1
/u/	336,7	332,8	3,9
Trung bình	**322,8**	**326,2**	**5,3**

Bảng 3.25 – Kết quả đo F0 với độ dài khung 20 ms của người nữ thứ nhất

Đơn vị đo: Hz

Tín hiệu	F0 đo thủ công	F0 tính tự động	Độ lệch của hai giá trị F0
/a/	199,6	230,3	30,8
/e/	184,5	224,1	39,5
/i/	226,2	228,2	2,0
/o/	179,3	225,9	46,6
/u/	239,7	243,9	4,2
Trung bình	**205,9**	**230,5**	**24,6**

Bảng 3.26 – Kết quả đo F0 với độ dài khung 20 ms của người nữ thứ hai

Đơn vị đo: Hz

Tín hiệu	F0 đo thủ công	F0 tính tự động	Độ lệch của hai giá trị F0
/a/	188,6	198,3	9,6
/e/	203,9	208,7	4,8
/i/	215,0	210,2	4,8
/o/	195,8	207,3	11,5
/u/	213,0	237,9	24,9
Trung bình	**203,3**	**212,5**	**11,1**

Bảng 3.27 – Kết quả đo F0 với độ dài khung 20 ms của người nữ thứ ba

Có thể thấy với độ dài khung 20 ms, độ lệch giữa cách tính F0 thủ công và cách tính F0 tự động bằng hàm tự tương quan đều giảm so với kết quả ở độ dài khung 15 ms ở cả 3 giọng nữ. Trong 3 giọng này thì người nữ thứ hai có sai số tính F0 lớn nhất (sai số tương đối xấp xỉ 12% với khung 20 ms và 18% với khung 15 ms), trong khi đó 2 người còn lại có sai số nhỏ hơn nhiều.

Với độ dài khung 30 ms, tôi thu được kết quả như sau:

Đơn vị đo: Hz

Tín hiệu	F0 đo thủ công	F0 tính tự động	Độ lệch của hai giá trị F0
/a/	315,1	319,3	4,3
/e/	310,7	318,1	7,4
/i/	334,1	332,0	2,1
/o/	317,4	318,5	1,1
/u/	336,7	332,6	4,1
Trung bình	**322,8**	**324,1**	**3,8**

Bảng 3.28 – Kết quả đo F0 với độ dài khung 30 ms của người nữ thứ nhất

Đơn vị đo: Hz

Tín hiệu	F0 đo thủ công	F0 tính tự động	Độ lệch của hai giá trị F0
/a/	199,6	255,4	55,9
/e/	184,5	233,4	48,9
/i/	226,2	227,6	1,4
/o/	179,3	227,9	48,7

/u/	239,7	238,2	1,5
Trung bình	**205,9**	**236,5**	**31,3**

Bảng 3.29 – Kết quả đo F0 với độ dài khung 30 ms của người nữ thứ hai

Đơn vị đo: Hz

Tín hiệu	F0 đo thủ công	F0 tính tự động	Độ lệch của hai giá trị F0
/a/	188,6	198,2	9,6
/e/	203,9	208,3	4,4
/i/	215,0	207,2	7,7
/o/	195,8	206,8	11,0
/u/	213,0	217,6	4,6
Trung bình	**203,3**	**207,6**	**7,5**

Bảng 3.30 – Kết quả đo F0 với độ dài khung 30 ms của người nữ thứ ba

So với các kết quả thu được ở độ dài khung 20 ms, kết quả tính F0 dùng độ dài khung 30 ms của người nữ thứ nhất và thứ ba có sai số giảm xuống (nhưng không đáng kể), trong khi đó kết quả của người nữ thứ hai lại có sai số tăng lên (sai số tương đối xấp xỉ 15% với khung 30 ms). Do đó không thể kết luận một cách rõ ràng độ dài khung 20 ms hay 30 ms cho kết quả tốt hơn với dữ liệu này.

Kết luận: từ các kết quả đo F0 tự động trên các tín hiệu của 3 người nữ đã khảo sát, độ chính xác của kết quả tính F0 có xu hướng tăng dần khi độ dài khung tín hiệu tăng từ 15 ms đến 30 ms. Thuật toán tự tương quan chưa cho thấy lỗi cao độ ảo khi áp dụng với 3 giọng nữ này.

3.9. Tổng kết chương

Trong chương này, tôi đã tiến hành cài đặt thuật toán tìm F0 dùng hàm tự tương quan của công cụ Matlab và dùng hàm tự tương quan tự viết, từ đó phát triển thành ứng dụng tính F0. Tôi đã khảo sát bằng thực nghiệm ảnh hưởng của các tham số quan trọng của các thuật toán để tìm ra bộ tham số tối ưu.

Đối với độ dài khung tín hiệu, độ chính xác của kết quả tính F0 có xu hướng tăng dần khi độ dài khung tín hiệu tăng từ 15 ms đến 30 ms cho cả giọng nam và nữ. Do đó, độ dài khung 30 ms nên được sử dụng để cho kết quả tính F0 đáng tin cậy nhất.

Với ngưỡng xác định hữu thanh/vô thanh, tôi đã tiến hành khảo sát và tìm ra ngưỡng biên độ của cực đại cục bộ bằng 30% giá trị biên độ của cực đại toàn cục của hàm tự tương quan cho kết quả chính xác nhất. Ngoài ra, tôi cũng tiến hành khảo sát giá trị kích thước của bộ lọc trung vị và xác định được với N = 7 thì bộ lọc trung vị cho kết quả làm trơn tối ưu.

KẾT LUẬN

1. Những việc đã hoàn thành

Với mục tiêu chính của đề tài là nghiên cứu phương pháp tính tần số cơ bản dựa trên hàm tự tương quan, sử dụng lọc trung vị để làm trơn kết quả và đánh giá ưu điểm và nhược điểm của các thuật toán, tôi đã thực hiện được các việc sau:

- Nghiên cứu lý thuyết liên quan đến xử lý tín hiệu tiếng nói, đặc biệt là tần số cơ bản F0 của tín hiệu tiếng nói.
- Nghiên cứu lý thuyết về hàm tự tương quan và thuật toán để tính F0 tự động từ tín hiệu tiếng nói.
- Nghiên cứu lý thuyết về lọc trung vị và áp dụng vào làm trơn chuỗi giá trị F0 đã tính được.
- Phân tích lý thuyết và khảo sát thực nghiệm các tham số quan trọng của các thuật toán xử lý tín hiệu.
- Cài đặt và đánh giá so sánh thuật toán tính F0 dùng hàm tự tương quan tự viết và hàm tự tương quan của Matlab, đánh giá độ chính xác của thuật toán tự viết trên dữ liệu thực nghiệm tự thu thập.

2. Các kết luận

Dựa trên các kết quả thực nghiệm, tôi đi đến các kết luận như sau:

- Hàm tự tương quan là phương pháp tương đối đơn giản và khá hiệu quả để giải quyết bài toán tính F0 của tín hiệu tiếng nói.
- Đối với cả giọng nam và nữ, thuật toán dùng hàm tự tương quan cho kết quả tính F0 chính xác nhất với độ dài khung tín hiệu dài 30 ms.
- Việc xác định một khung tín hiệu là tuần hoàn hay không tuần hoàn (thuộc về âm hữu thanh hay âm vô thanh) dựa trên hàm tự tương quan của tín hiệu bị ảnh hưởng bởi ngưỡng biên độ của cực đại cục bộ. Khảo sát định tính cho thấy ngưỡng này bằng 30% giá trị biên độ của cực đại toàn cục của hàm tự tương quan thì cho kết quả chính xác nhất.
- Khi sử dụng hàm tự tương quan để tính F0, cần phải có bước lọc trung vị trên kết quả F0 nhận được để có đường F0 biến đổi trơn tru hơn. Kích thước của bộ lọc trung vị với N = 7 thì bộ lọc trung vị cho kết quả làm trơn tối ưu.

- Thuật toán tính F0 dùng hàm tự tương quan khá nhạy với lỗi độ cao ảo. Cần có các giải pháp để khắc phục vì lỗi này tạo ra sai số rất lớn.

3. Hạn chế và hướng phát triển

Do thiếu dữ liệu F0 chuẩn cho từng khung tín hiệu tiếng nói nên việc đánh giá độ chính xác của thuật toán tính F0 chỉ dùng một thước đo sai số tương đối đơn giản (độ lệch của giá trị F0 giữa cách đo thủ công và thuật toán dùng hàm tự tương quan, tính cho cả tín hiệu) là chưa đầy đủ. Các tín hiệu dùng trong thực nghiệm mới chỉ dừng ở các nguyên âm, vốn có đặc trưng tuần hoàn tương đối ổn định, nên chưa đánh giá được đầy đủ hiệu năng của thuật toán khi gặp tín hiệu tiếng nói có tính chất biến đổi phức tạp hơn. Luận văn cũng chưa đưa ra các cải tiến để cải thiện độ chính xác (đặc biệt là khắc phục lỗi cao độ ảo) và tốc độ thực thi của thuật toán, và thiếu đánh giá so sánh với các thuật toán tính F0 khác. Tìm giải pháp cho các vấn đề trên là hướng phát triển của luận văn trong tương lai.

TÀI LIỆU THAM KHẢO

[1] Phạm Văn Sự, Lê Xuân Thành, *Bài giảng Xử lý tiếng nói*, Học viện Công nghệ Bưu chính Viễn thông, 2010.

[2] Nguyễn Bình Thiên, Ninh Khánh Duy, *Cải tiến thuật toán tự tương quan tìm cao độ của tín hiệu đàn ghi-ta trên vi xử lý ARM Cortex-M4*, Kỷ yếu hội nghị quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), 2017.

[3] Lê Tiến Thường, Huỳnh Ngọc Phiên, Trần Tiến Đức, *Phương pháp mới trích chu kỳ cao độ trung bình ứng dụng trong nhận dạng thanh điệu tiếng Việt*, Tạp chí Bưu chính Viễn thông và Công nghệ thông tin, 2003.

[4] Lawrence R. Rabiner and Ronal W.Schafer , *Digital Processing Of Speech Signals*, Prentice Hall, 1978.

[5] John G. Proakis and Dimitris G. Manolakis, *Digital Signal Processing: Principles, Algorithms & Applications*, Prentice Hall, 1995.

[6] Vinay K. Ingle and John G. Proakis, *Digital Signal Processing Using MATLAB*, Cengage Learning, 2012.

[7] Denis Jouvet and Yves Laprie, *Performance Analysis of Several Pitch Detection Algorithms on Simulated and Real Noisy Speech Data*, 25th European Signal Processing Conference (EUSIPCO), 2017.

[8] Alain de Cheveigne and Hideki Kawahara, *Comparative evaluation of F0 estimation algorithms*, Eurospeech, 2001.

[9] Duy Khanh Ninh and Yoichi Yamashita, *F0 Parameterization of Glottalized Tones in HMM-Based Speech Synthesis for Hanoi Vietnamese*, IEICE TRANS. INF. & SYST., 2015.