# Mô hình Fujisaki và áp dụng trong phân tích thanh điệu tiếng Việt

Bạch Hưng Nguyên, Nguyễn Tiến Dũng
Viện Công Nghệ Thông Tin
Trung Tâm Khoa Học Tự Nhiên & Công Nghệ Quốc Gia
nguyenbh@netnam.org.vn, nguyentiendung@hotmail.com

## Tóm tắt
Trong bài báo này chúng tôi trình bầy những nghiên cứu bước đầu về việc áp dụng mô hình Fujisaki cho tổng hợp tiếng Việt có ngữ điệu. Các câu nói được thiết kế để vừa mang đủ sáu thanh điệu vừa thể hiện các tổ hợp thanh quan trọng như thanh ngã và thanh nặng. Tham số mô hình đã được điều chỉnh để thích ứng với các đặc trưng của ngôn ngữ tiếng Việt.

## 1. Giới thiệu

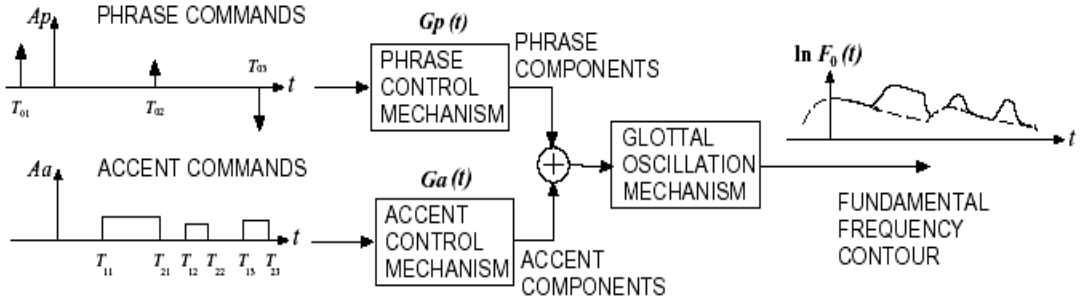
Giải thích một cách nôm na ngôn điệu chính là cái mang lại cho tiếng nói con người những âm sắc riêng biệt. Nếu một đoạn tiếng nói mà không chứa ngôn điệu thì nó giống như giọng nói của người máy, không giống tiếng nói tự nhiên. Các nhà ngôn ngữ học cho rằng bản chất ngôn điệu là các hiện tượng phủ lên âm tiết trọn vẹn chẳng hạn như trọng âm, thanh điệu, và ngữ điệu; ngoài ra còn có các hiện tượng bên trong âm tiết nhưng không thể qui cho từng chiết đoạn bộ phận mà âm tiết bao hàm; hiên tượng thứ ba là trường độ. Vai trò ngôn điệu rất quan trọng trong tổng hợp tiếng nói. Nếu không xử lý được vấn đề ngôn điệu thì không thể có được tiếng nói tổng hợp giống tiếng nói tự nhiên. Các đặc trưng quan trọng nhất của ngôn điệu là độ cao, độ dài, và độ to, tương ứng là các đại lượng tần số cơ bản F0, thời gian của âm tiết, âm vị D, và cường độ I.

Ngôn điệu của lời nói liên kết chặt chẽ với khái niệm "ngữ điệu". Có thể nói ngữ điệu là sự nâng cao hạ thấp của giọng nói trong câu. Tần số cơ bản F0 là đặc trưng chính của ngữ điệu. Ngữ điệu là một thành phần của ngôn điệu. Tiếng Việt là ngôn ngữ có thanh điệu, các thanh điệu có các đặc trưng rất khác nhau về đường nét F0. Trong lời nói liên tục, đường nét F0 của các thanh điệu bị biến đổi phụ thuộc vào thanh điệu của các âm tiết liền kề và vị trí của âm tiết trong câu. Việc mô hình hoá đường nét F0 các thanh điệu có ý nghĩa quan trọng trong việc tổng hợp tiếng nói.

Fujisaki và các đồng sự đã phát triển một cách mô tả toàn diện ngữ điệu tiếng Nhật dựa trên một mô hình định lượng sau này mang tên Fujisaki [2]. Mô hình Fujisaki được ứng dụng rộng rãi trong các hệ thống tổng hợp của tiếng Nhật như tổng hợp các bản tin thời tiết. Mô hình MFGI (Mixdorff-Fujisaki model of German Intonation) được ứng dụng trong hệ thống Text-to-Speech tiếng Đức. Với một số thay đổi nhỏ, mô hình Fujisaki thích hợp trong việc phân tích đường nét F0 trong tiếng Anh, tiếng Thụy Điển, tiếng Tây Ban Nha, tiếng Đức, tiếng Hy Lạp, phân tích và tổng hợp thanh điệu của ngôn ngữ có thanh điệu như tiếng Trung, tiếng Thái.

Trong bài báo này chúng tôi điều chỉnh việc áp dụng mô hình cho tiếng Đức, tiếng Trung, và tiếng Thái, đồng thời tiến hành thử nghiệm mô hình Fujisaki với các câu nói tiếng Việt. Các câu nói được thiết kế để vừa mang đủ sáu thanh điệu vừa thể hiện các tổ hợp thanh quan trọng như thanh ngã và thanh nặng. Các tham số mô hình đã được điều chỉnh để thích ứng với các đặc trưng của ngôn ngữ tiếng Việt.

## 2. Mô hình Fujisaki




Hình 1: Mô hình Fujisaki

Fujisaki là một mô hình định lượng dùng để mô hình hóa ngữ điệu (intonation). Mô hình Fujisaki hướng vào việc mô hình hóa quá trình sinh ra tần số cơ bản F0, giải thích về mặt vật lý học, sinh lý học quá trình sinh ra F0 và các tính chất của quá trình đó. Mô hình được áp dụng chủ yếu trong ứng dụng tổng hợp nhằm xây dựng phần ngữ điệu trong tiếng nói tổng hợp. Mô hình sinh ra F0 theo bộ ba công thức sau [2]:

$$\ln F0(t) = \ln Fb + \sum_{i=1}^{I} Ap_i Gp(t - T_{0i}) + \sum_{j=1}^{J} Aa_j \left[ Ga(t - T_{1j}) - Ga(t - T_{2j}) \right] \quad (1)$$

$$Gp(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & t \geq 0, \\ 0, & t < 0 \end{cases} \quad (2)$$

$$Ga(t) = \begin{cases} \min\left[1 - (1 + \beta t)\exp(-\beta t), \gamma\right], & t \geq 0, \\ 0, & t < 0 \end{cases} \quad (3)$$

Các tham số của mô hình gồm có

- Các hằng số: Fb là giá trị khởi đầu của đường tần số cơ bản. Fb là giá trị phụ thuộc vào người nói chứ không phụ thuộc vào các mẫu tiếng nói. Giá trị α là tần số góc tự nhiên của *lệnh ngữ* (phrase command). Giá trị β là tần số góc tự nhiên của *lệnh trọng âm* (accent command). Giá trị γ là mức giá trị trần tương ứng với các thành phần trọng âm
- Các đối số: I là số lệnh ngữ. J là số lệnh trọng âm. $Ap_i$ là cường độ của lệnh ngữ thứ i. $Aa_j$ là biên độ của lệnh trọng âm thứ j. $T_{0i}$ là thời điểm bắt đầu lệnh ngữ thứ i. $T_{1j}$ và $T_{2j}$ là thời điểm bắt đầu và kết thúc thanh điệu ở lệnh trọng âm thứ j.

Trong mô hình, đường F0 được xét ở miền logF0, mục đích của phép biến đổi này là làm cho giọng nói của nam và nữ giống nhau. Theo [1] các giá trị α =2.0/s và β =20.0/s, trong một số trường hợp đặc biệt α =3.0/s. Tuy nhiên theo quan sát thì α nằm trong khoảng [1.0;3.0], còn β thuộc khoảng [19.5;20.5].

**Thành phần ngữ** Gp(t) trong công thức (2) định nghĩa cơ chế điều khiển ngữ. Đầu vào của cơ chế điều khiển ngữ là các lệnh ngữ bao gồm cường độ Ap với thời gian bắt đầu T0. Hệ số α là hằng số thời gian và là không đổi với một câu nói.

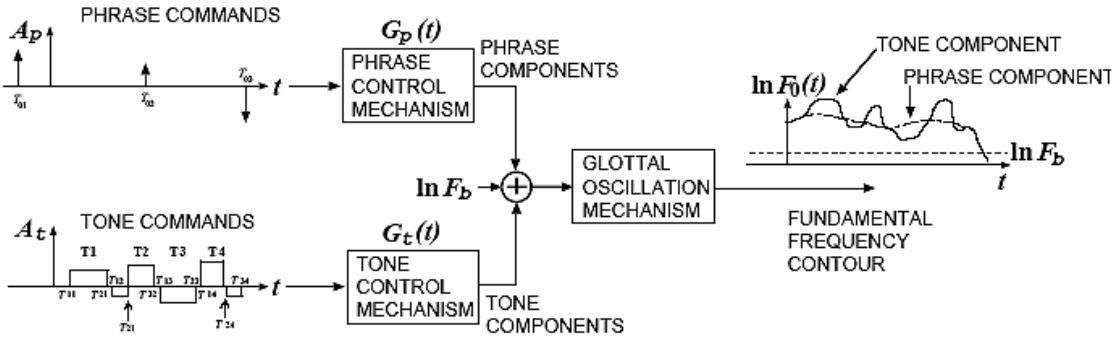
**Thành phần trọng âm** Ga(t) trong công thức (3) định nghĩa cơ chế điều khiển trọng âm với đầu vào là các lệnh trọng âm bao gồm biên độ Aa, thời gian bắt đầu T1, thời gian kết thúc T2. Hệ số β là hằng số theo thời gian của cơ chế điều khiển trọng âm và là không đổi với một câu nói. Thành phần trọng âm không bao giờ vượt quá giá trị trần γ (thường được gán giá trị 0.9).

Việc phân tích đường nét F0 được thực hiện bởi phương pháp **phân tích bằng tổng hợp** (viết tắt là AbS: Analysis-by-Synthesis). Giá trị các tham số của mô hình được thay đổi cho tới khi xấp xỉ tốt nhất đường nét F0 của câu nói được phân tích. Với số lượng lệnh không giới hạn

(lệnh ngữ và lệnh trọng âm), bất kỳ đường nét F0 nào cũng có thể được xấp xỉ với độ chính xác không giới hạn. Vì thế cần có các ràng buộc để đảm bảo tính có nghĩa về mặt ngôn ngữ học của các kết quả phân tích. Các ràng buộc đó là các đặc trưng về ngôn ngữ và liên quan tới mối quan hệ giữa các cấu trúc và đơn vị ngôn ngữ (như ngôn điệu và trọng âm) và các lệnh ngữ và lệnh trọng âm.

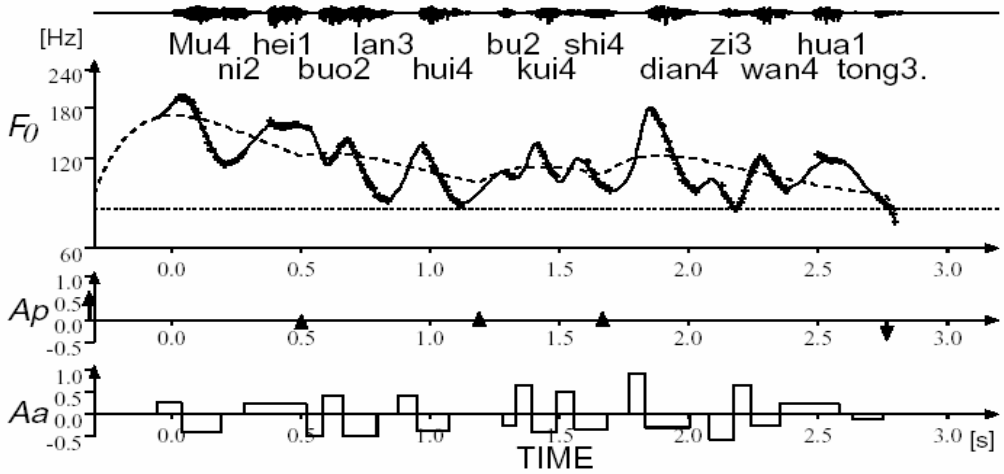
Các tham số Ap, T0, α, Aa, T1, T2, β, Fb được gọi là là các tham số Fujisaki và phương pháp phân tích bằng tổng hợp đường nét F0 sử dụng mô hình Fujisaki được gọi là phân tích Fujisaki. Các tham số của mô hình có thể được sinh ra tự động bởi nhiều cách khác nhau tùy vào từng ngôn ngữ được phân tích [8].

### 2.1 Ứng dụng của mô hình Fujisaki cho ngôn ngữ có thanh điệu




Hình 2: Mô hình Fujisaki khi áp dụng cho các ngôn ngữ có thanh điệu
có thêm các lệnh thanh điệu âm

Khi áp dụng mô hình Fujisaki cho các ngôn ngữ có thanh điệu, thành phần trọng âm được gọi là thành phần thanh điệu và sử dụng cả lệnh thanh điệu dương ($A_t > 0$), lệnh thanh điệu âm ($A_t < 0$) để mô hình hoá các thanh điệu như trong hình 2.




Hình 3: Áp dụng mô hình Fujiaki cho phân tích một câu tiếng Trung

Các âm tiết tiếng Trung có 4 thanh điệu [4]: T1: high tone, T2: rising tone, T3: low tone, T4: falling tone. Mặc dù đường nét F0 của các thanh điệu rất khác nhau, nhưng chúng vẫn bị thay đổi đáng kể lời nói liên tục bởi những yếu tố như thanh điệu của những âm tiết liền kề, cú pháp.

Các thanh điệu tiếng Trung được mô hình hóa bởi mô hình Fujisaki như sau: T1 và T3 được tạo ra bởi một lệnh thanh điệu, lệnh thanh điệu dương cho T1 và lệnh thanh điệu âm cho T3.

T2 và T4 được tạo ra bởi một cặp lệnh thanh điệu liền nhau, cặp lệnh thanh điệu âm-dương cho T2 và cặp lệnh thanh điệu dương-âm cho T4.

Kết quả phân tích một số câu nói tiếng Trung cho thấy mô hình luôn xấp xỉ rất tốt đường nét F0. Những kết quả phân tích có thể dùng làm luật để sinh ra khoảng thời gian của lệnh thanh điệu trong tổng hợp tiếng nói.

### 2.2 Nhận xét về mô hình Fujisaki

Đây là mô hình duy nhất đưa vào nền tảng vật lý học và sinh lý học của quá trình sinh ra F0. Thêm vào đó là mô hình duy nhất có các công thức toán học sinh ra được đường nét F0 bất kỳ, cho phép xác định số lượng của các sự kiện ngữ điệu. Các sự kiện ngữ điệu được gắn với các mốc thời gian rõ ràng. Hơn nữa việc tổng hợp F0 là dễ dàng. Đường nét F0 liên tục được phân tích thành các phần đơn vị ngữ điệu rời rạc (các lệnh) với biên độ liên tục. Ngoài ra, đường nét F0 có thể mô hình hóa với độ chính xác cao với một số lượng nhỏ các tham số. Cuối cùng trong quá trình mô hình hóa, đường nét F0 sinh ra được làm trơn và bỏ đi những biến đổi rất nhỏ về ngôn điệu.

Mô hình Fujisaki sinh ra đường nét F0 đã được làm trơn nên tiếng nói tổng hợp sử dụng mô hình Fujisaki nghe mềm mại và thật hơn so với các phương pháp mô hình hóa đường nét F0 khác. Tuy nhiên việc xác định các tham số của mô hình bằng phương pháp phân tích bằng tổng hợp (Analysis-by-Synthesis) đòi hỏi người phân tích phải có kinh nghiệm và kiên trì. Việc xác định các hệ số α, β là khó, hầu như phụ thuộc vào việc kiểm tra lại và tối ưu hóa dần.

Mô hình Fujisaki đã được áp dụng thành công trong việc mô hình hóa đường nét F0 của các ngôn ngữ có thanh điệu, đặc biệt là tiếng Trung cho thấy mô hình có thể áp dụng được trong việc mô hình hóa thanh điệu tiếng Việt.

## 3. Phân tích thanh điệu tiếng Việt bằng mô hình Fujisaki

### 3.1 Cơ sở dữ liệu

Để phân tích đường nét F0 của thanh điệu tiếng Việt và sự liên cấu âm giữa các thanh điệu liền kề, một tập gồm 72 câu nói, mỗi câu nói gồm 6 âm tiết được xây dựng từ câu gốc: "nha mai lăm nhan nhiêu ngô", mỗi âm tiết trong câu gốc sẽ mang các thanh điệu khác nhau để thể hiện được nhiều tổ hợp thanh điệu liền kề như:

1) Nhà mai lắm nhãn nhiều ngô
2) Nhà mài lắm nhan nhiêu ngộ
3) Nha mải lắm nhãn nhiêu ngổ

…….....

Các câu được phát âm với giọng chuẩn miền Bắc bởi hai người một nam, một nữ. Để thể hiện được nhiều tổ hợp âm chỉ dùng một câu gốc nên đa số các câu trong cơ sở dữ liệu đều không có nghĩa. Để đảm bảo tính tự nhiên của lời nói, hai người nói đều được chuẩn bị trước, các câu nói được phát âm nhiều lần và kiểm tra lại để chọn câu nói tự nhiên nhất. Đường nét F0 được tính toán theo từng đoạn 10ms.

### 3.2 Phương pháp phân tích

Để phân tích đường nét F0, một công cụ phân tích các tham số của mô hình Fujisaki được sử dụng. Công cụ này hỗ trợ các lệnh thanh điệu âm và đã được sử dụng trong phân tích đường nét F0 tiếng Thái [6]. Fb được đặt bằng 96 Hz cho giọng nam và 210 Hz cho giọng nữ. α và β cho cả giọng nam và nữ được lần lượt đặt bằng 2 Hz và 25 Hz.

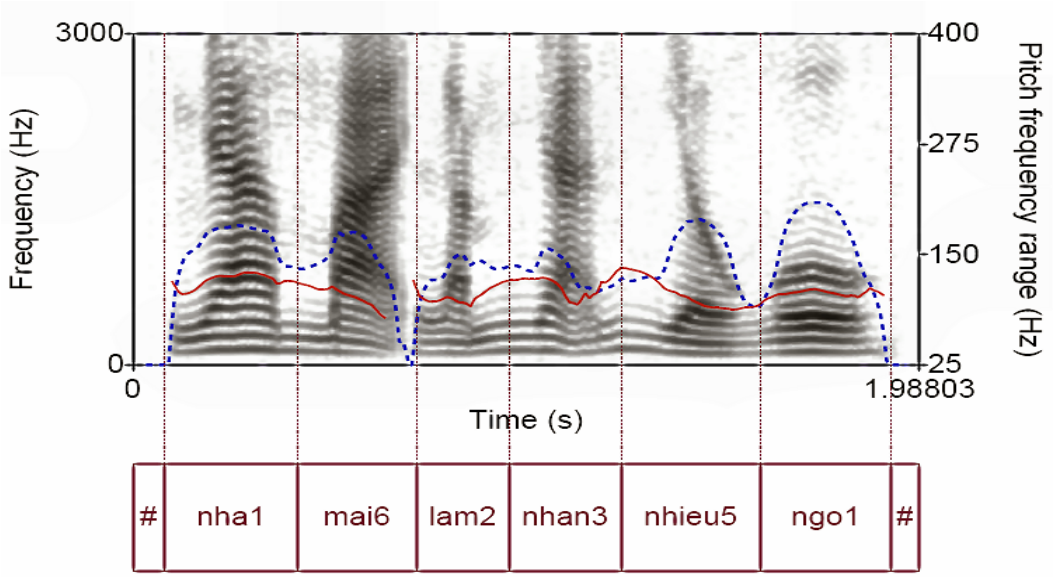
Các bước tiến hành phân tích bao gồm:

1. Tính đường nét F0.
2. Lựa các chọn lệnh ngữ câu nói.

3. Dựa vào thanh điệu của các âm tiết để lựa chọn các lệnh thanh điệu phù hợp.
4. Điều chỉnh các tham số sao cho đường nét F0 sinh ra xấp xỉ tốt đường nét F0 thực.
5. Tổng hợp lại câu nói với đường nét thanh điệu mới sử dụng phương pháp PSOLA.
6. Cảm nhận bằng tai câu nói tổng hợp, so sánh với câu nói gốc và điều chỉnh lại.

### 3.3 Một số nhận xét về thanh điệu tiếng Việt trong khi phân tích cơ sở dữ liệu

**Sự khác biệt giữa thanh nặng và ngã với các thanh điệu khác** là rõ ràng. Trong 6 thanh điệu tiếng Việt, thanh nặng và thanh ngã có những đặc điểm khác biệt so với các thanh điệu còn lại. Đường nét F0 bị đứt đột ngột ở hai thanh này. Các thanh này không chỉ khác các thanh điệu khác ở đường nét F0 mà còn ở các đặc trưng khác. Do đó, khi cần tổng hợp lại thanh ngã và thanh nặng thì chỉ đường nét F0 là chưa đủ.




Hình 4: Phổ, đường nét F0 (đường màu đậm) và cường độ (đường nét đứt) của trong câu "nha mại lắm nhãn nhiều ngô".

**Thanh ngã** bị gãy ở giữa. Không những bị gãy ở F0, thanh ngã còn bị gãy ở phổ, cường độ đó chính là khác biệt lớn nhất giữa thanh ngã và các thanh khác.

**Thanh nặng** có đặc trưng là bị gẫy, đứt, và đi xuống đột ngột ở cuối âm, lúc này đường nét F0 không còn quan trọng, ví dụ như khi cho thanh nặng đường nét F0 của thanh sắc thì người nghe vẫn cảm nhận được thanh nặng và chỉ có cảm tưởng người nói nhấn mạnh hơn ở âm tiết mang thanh nặng.

Ở những âm tiết đóng kết thúc bằng t, p, c, k (chỉ có thể có 2 thanh sắc hoặc nặng), khi cho đường nét của thanh nặng đi lên rồi tổng hợp lại ta nhận được thanh sắc và ngược lại, cho đường nét thanh sắc đi xuống rồi tổng hợp lại ta nhận được thanh nặng. Ở các âm tiết khác điều này là không đúng.

**Sự giống nhau giữa thanh ngã và thanh sắc** đó là cả hai thanh đều có đường nét thanh điệu đi lên, và âm vực bắt đầu của âm sắc cao hơn thanh ngã. Khi làm thí nghiệm có hiện tượng: Cho đường nét của thanh sắc và ngã giống hệt nhau, khi tổng hợp lại người nghe vẫn phân biệt được hai thanh này.

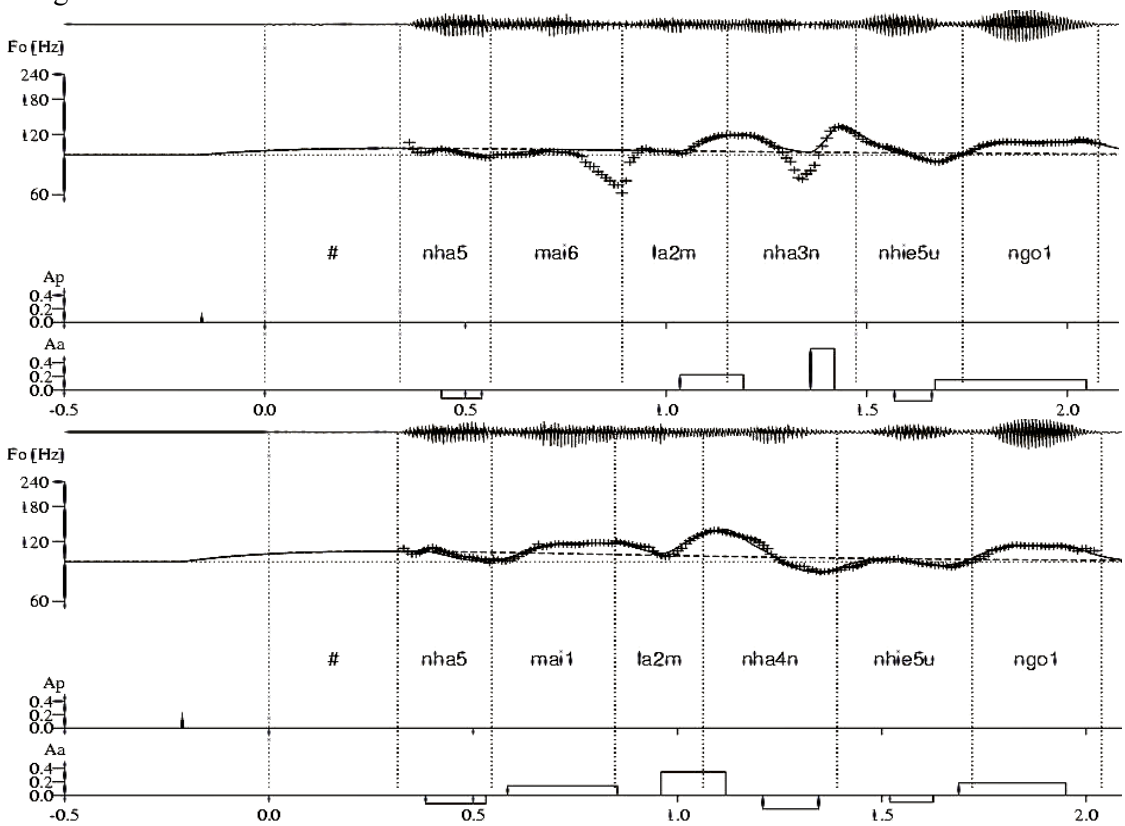
### 3.4 Kết quả phân tích thanh điệu bằng mô hình Fujisaki [9]

Các kết quả phân tích cơ sở dữ liệu cho thấy, các thanh ngang, sắc, ngã được biểu diễn bằng một lệnh thanh điệu dương, thanh huyền và hỏi được biểu diễn bằng một lệnh thanh điệu âm, thanh nặng không cần lệnh thanh điệu.

| Thanh điệu | Biểu diễn bằng lệnh thanh điệu |
|---|---|
| Ngang | 1 lệnh thanh điệu dương ở trước âm tiết |
| Sắc | 1 lệnh thanh điệu dương |
| Hỏi | 1 lệnh thanh điệu âm |
| Huyền | 1 lệnh thanh điệu âm |
| Ngã | 1 lệnh thanh điệu dương |
| Nặng | không dùng lệnh thanh điệu |

Bảng 1: Biểu diễn các 6 thanh điệu tiếng Việt bằng các lệnh thanh điệu

Các câu được phân tích chỉ sử dụng một lệnh ngữ cho cả câu, phù hợp với hiện tượng trong câu nói, người nói thường lên giọng ở đầu câu và hạ giọng ở cuối câu. Tuy nhiên trong tiếng Việt hiện tượng này không rõ rệt như ở các ngôn ngữ khác nên cường độ của lệnh ngữ này không lớn.




Hình 5: Kết quả phân tích thanh điệu tiếng Việt bằng mô hình Fujisaki

Thanh ngã và thanh sắc được biểu diễn bằng môt lệnh ngữ điệu dương phù hợp với nhận xét về sự giống nhau giữa 2 thanh này trong phần trước.

Thanh hỏi có đường nét F0 đi xuống, đến giữa thanh, đường nét F0 lại đi lên, thanh này giống thanh T3 (low tone) của tiếng Trung và được biểu diễn bằng một lệnh thanh điệu âm giống như trường hợp của tiếng Trung.

## 4. Kết luận

Mô hình về cơ bản không thể áp dụng cho bài toán nhận dạng tiếng nói. Lí do chủ yếu là mô hình này thực chất tổng hợp đường F0 một cách tuyến tính, nó là một phép biến đổi làm trơn đường F0. Đầu vào của mô hình này là một F0 thô, khi đi qua mô hình ta được một F0 mới với các đặc trưng về ngữ điệu, hay còn gọi là F0 trơn. Các kết quả phân tích thanh điệu tiếng Việt chứng tỏ rằng có thể áp dụng mô hình Fujisaki vào việc mô hình hóa thanh điệu tiếng Việt, cao hơn nữa là mô hình hóa ngữ điệu tiếng Việt với việc thể hiện đường nét F0 không chỉ của các thanh điệu mà còn của cả các loại câu như câu trần thuật, câu hỏi, ... Từ đó làm nâng cao chất lượng của hệ thống tổng hợp tiếng nói và các kết quả phân tích cũng có thể áp dụng kết quả tính toán ngữ âm học vào việc nhận dạng tiếng nói. Tuy nhiên, để có thể làm được điều trên cần phải có những phân tích sâu hơn nữa trên bộ dữ liệu thực lớn hơn, phân tích thanh điệu dưới ngữ cảnh của các thanh điệu khác.

## Tài liệu tham khảo


[1] H. Mixdorff, "Intonation patterns of German model based quantitative analysis and synthesis of F0 contour", PhD Thesis, TFH Berlin University.

[2] H. Fujisaki, S. Ohno, C. Wang, "A command-response model for F0 contour generation in multilingual speech synthesis", Journal of Phonetics, vol. 2, pp 223-232, 1974.

[3] Q.C. Nguyen, E. Castelli, N.Y. Pham, "Tone recognition for Vietnamese", Automatic Speech Recognition and Understanding workshop 2001.

[4] C. Wang, H. Fujiaski, K. Hirose, "The four tones recognition of continuous Chinese speech", International Conference on Spoken Language Processing, pp. 221-224, 1990.

[5] C. Wang, H. Fujisaki, S. Ohno, T. Kodama, "Analysis and synthesis of the four tones in connected speech of the Standard Chinese based on a command-response model", Proceeding of the 6th European Conference on Speech Communication and Technology, vol. 4, pp. 1655-1658, 1999.

[6] N. Thubthong, A. Pusittrakul, T. Sookawat, B. Kijsirikul, "Tone recognition of continuous Thai using half-tone model", National Computer Science and Engineering Conference (NCSEC'2000).

[7] H. Mixdorff, S. Luksaneeyanawin, H. Fujisaki, et al, "Perception of tone and vowel quantity in Thai", International Conference on Spoken Language Processing 2002 at Denver, USA.

[8] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters", Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing 2000, pp.1281-1284.

[9] N. H. Bach, H. Mixdorff, H. Fujisaki, M. C. Luong, "Quantitative analysis and synthesis of syllabic tones in Vietnamese", Proceeding of the 8th European Conference on Speech Communication and Technology, 2003.