Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00092

PHƯƠNG PHÁP HIỆU QUẢ XÁC ĐỊNH CÁC ĐIỂM ĐÁNH DẤU PITCH VÀ HỒI QUY PHI TUYẾN ĐỂ TỔNG HỢP THANH ĐIỆU CỦA ÂM TIẾT TIẾNG VIỆT RỜI RẠC

Tạ Yên Thái[1], Nguyễn Văn Hùng[2], Ngô Hoàng Huy[3], Phạm Kim Thư[1]

[1] Đại học Kinh doanh và Công nghệ Hà Nội: Số 29A, Ngõ 124, Phố Vĩnh Tuy, Q. Hai Bà Trưng, Hà Nội.

[2] Viện Công nghệ thông tin, Viện Khoa học và Công nghệ Quân sự: Số 17 Hoàng Sâm, Q. Cầu Giấy, Hà Nội

[3] Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam:
Số 18 Hoàng Quốc Việt, Q.Cầu Giấy, Hà Nội

{tayenthai, nvhnt73, huyngo3i, thuphamkim}@gmail.com

***TÓM TẮT**: Trong vấn đề nghiên cứu về tổng hợp tiếng Việt, tổng hợp thanh điệu đóng một vai trò quan trọng và thể hiện tính đặc thù của tiếng Việt. Các thuật toán tổng hợp thanh điệu tiếng Việt thường sử dụng phương pháp bình phương tối thiểu để cách điệu hóa tuyến tính đường F0 của các thanh điệu tiếng Việt phát âm rời rạc hoặc liên tục trong ngữ lưu. Tiếp cận theo hướng này là rất khó để cách điệu hóa tuyến tính đường F0 của một số thanh điệu tiếng Việt như thanh nặng và thanh ngã.*

Bên cạnh đó việc ước lượng hiệu quả các điểm đánh dấu pitch (PM) của một phát âm là bước quan trọng đầu tiên để tổng hợp thanh điệu cho tới nay vẫn là một vấn đề mở. Bài báo này trình bày một phương pháp ước lượng các PM của sóng tiếng nói và phương pháp xây dựng một phép hồi quy phi tuyến để tổng hợp các thanh điệu dựa trên mô hình Xu đã được sử dụng rộng rãi cho tiếng Trung Quốc phổ thông –Mandarin. Thực nghiệm đã chứng tỏ tính hiệu quả của thuật toán đề xuất cho vấn đề xác định các điểm đánh dấu pitch của tín hiệu tiếng nói đầu vào dùng cho tổng hợp thanh điệu tiếng Việt phát âm rời rạc. Các điểm đánh dấu pitch được ước lượng với độ chính xác cao. Các thanh điệu tổng hợp nghe rõ ràng cả với thanh nặng và thanh ngã, giữ được đường nét đặc trưng thanh điệu tương ứng.

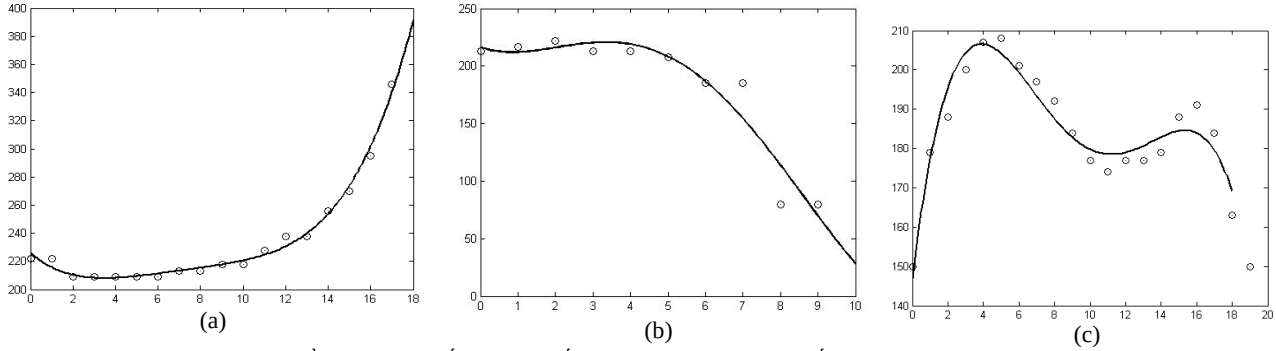
***Từ khóa**: Hữu thanh/vô thanh, Cách điệu đường F0, mô hình Xu, Điểm đánh dấu pitch, Tổng hợp thanh điệu.*

I. GIỚI THIỆU

Mỗi âm tiết tiếng Việt nhất thiết phải được thể hiện với một thanh điệu. Thanh điệu có chức năng phân biệt vỏ âm thanh, phân biệt nghĩa của từ. Nếu như ngữ điệu là đặc trưng của câu, trọng âm là đặc trưng của từ thì thanh điệu là đặc trưng của âm tiết tiếng Việt.

Sáu thanh điệu tiếng Việt được chia thành hai nhóm lớn bằng và trắc. Thanh không dấu và thanh huyền thuộc loại thanh bằng có đường nét tương đối đơn giản. Thanh ngã, thanh hỏi, thanh sắc và thanh nặng là những thanh trắc có đường nét thanh điệu phức tạp. Các thanh ngang, sắc, ngã thuộc âm vực cao, còn các âm huyền, hỏi và nặng thuộc âm vực thấp. Ở mức vật lý, phần thanh của thanh điệu chính là đường nét của tần số âm cơ bản F0.

Tần số cơ bản F0 mang tính tương đối, đặc trưng cho từng thanh điệu. Đường nét F0 được xác định bằng sự biến đổi tần số dao động của dây thanh, do các cơ thanh quản, cũng như áp suất dòng khí đi qua thanh môn điều phối.



Hình 1. Các ví dụ về dạng thanh sắc của âm tiết thử nghiệm (a) Thanh sắc, (b) Thanh nặng và (c) Thanh ngã

Do bản chất của tiếng Việt, vấn đề tổng hợp thanh điệu tiếng Việt đã được quan tâm nghiên cứu [1-6]. Trong các nghiên cứu này các tác giả đã sử dụng các mô hình cách điệu đường F0 khác nhau để tổng hợp thanh điệu âm tiết rời như [1-2] với cách điệu tuyến tính, [3-5] với phương pháp phân tích và tổng hợp sử dụng mô hình Fujisaki hoặc dựa trên học thống kê với mô hình học máy HMM [6].

Do hiện tượng biến thanh của thanh điệu khá phức tạp nên các phương pháp tổng hợp thanh điệu cho câu tự nhiên sử dụng mô hình Fujjisaki hoặc HMM vẫn là vấn đề mở. Ngược lại, hình dáng thanh điệu của âm tiết hoặc từ tiếng Việt phát âm rời khá ổn định, có nhu cầu ứng dụng trong thực tế như huấn luyện phát âm thanh điệu tiếng Việt, hỗ trợ người khiếm thính sử dụng máy tính (phát âm ký tự/ từng âm tiết được gõ) v.v... Cách điệu hóa đường F0 của

âm tiết phát âm rời để tổng hợp thanh điệu yêu cầu đòi hỏi một phương pháp hiệu quả thay thế mô hình Fujisaki (có nhiều tham số, yêu cầu giai đoạn phân tích tốn nhiều thời gian và công sức) hoặc mô hình thống kê HMM không thích hợp cho xử lý mức đoạn âm ngắn mức âm tiết/từ. Phương pháp cách điệu tuyến tính đơn giản hơn, tuy vậy phương pháp này lại khó tổng hợp được các âm tiết mang thanh nặng và thanh ngã từ âm tiết không dấu [1].

Với các ngôn ngữ có thanh điệu như tiếng Trung quốc phổ thông Mandarin, tiếng Thái mô hình "pitch target" của Xu và cộng sự được nghiên cứu và sử dụng hiệu quả để sinh đường F0 mức âm tiết/từ/ngữ đoạn, xem chẳng hạn [7-12]. Các tham số mô hình được khớp (fitting) bằng gói phần mềm SPSS. Ưu điểm của mô hình là đơn giản, ít tham số và có thể học thống kê để sinh ra đường biểu diễn F0 trong ngữ cảnh. Đầu vào của mô hình là các giá trị F0 được tính bằng thuật toán ước lượng F0 gọi là RAPT (Robust Algorithm for Pitch Tracking) của David Talkin [13] trong chương trình get_f0 của gói ESPS (Entropic Signal Processing System) [19]. Cho đến nay vấn đề ước lượng giá trị F0 vẫn là vấn đề mở, có rất nhiều thuật toán để ước lượng F0, xem chẳng hạn [14, 15] và đích ứng dụng của các thuật toán này thường là nhận dạng và xử lý tiếng nói trong môi trường thực. Tuy nhiên trong vấn đề tổng hợp đường F0, tín hiệu tiếng nói có chất lượng cao, được thu nhận trong môi trường phòng thí nghiệm, không có nhiễu mạnh. Do vậy chúng ta cần một phương pháp riêng, hiệu quả để ước lượng F0 làm đầu vào cho bộ cách điệu F0 để tổng hợp thanh điệu cho âm tiết/từ,…

Để tổng hợp thanh điệu của âm tiết tiếng Việt phát âm rời, trong bài báo này chúng tôi trình bày thuật toán ước lượng F0 cho đoạn tiếng nói hữu thanh (như đoạn phần vần của tín hiệu âm tiết tiếng Việt), phép giải ước lượng hệ số hồi quy phi tuyến của mô hình Xu. Chi tiết việc các thuật này được trình bày trong phần III và phần IV của bài báo.

Phần còn lại của bài báo được tổ chức như sau: Phần II, trình bày một số nghiên cứu liên quan của thuật toán Xu; Phần III là đề xuất thuật toán ước lượng F0, thuật toán ước lượng hệ số hồi quy phi tuyến của mô hình Xu; Các kết quả thực nghiệm đưa ra trong phần IV; Kết luận được cho trong phần V.

II. NGHIÊN CỨU LIÊN QUAN

Để biến đổi đường F0 của một ngữ đoạn âm thanh, âm tiết thành một ngữ đoạn có đường F0 thay đổi, âm tiết mang thanh điệu khác, chúng ta cần xác định dạng đường F0 đích và các điểm đánh dấu pitch (PM-Pitch mark, hoặc các Pulse) của âm thanh gốc [16, 17], sau đó sử dụng một công cụ nắn chỉnh đường F0 chẳng hạn thuật toán PSOLA [1]. Do vậy việc xác định giá trị đường F0 của ngữ đoạn tiếng nói, các điểm PM và cách điệu nó để khái quát cho đường F0 âm thanh đích là điểm quan trọng để tổng hợp tiếng Việt [1].

A. Ước lượng F0 và các điểm đánh dấu pitch của đoạn tín hiệu tiếng nói hữu thanh

Một cách ước lượng F0 đơn giản là sử dụng phương pháp tự tương quan. Chia đoạn tín hiệu tiếng nói thành các khung thời gian ngắn (gọi là frame) có độ dài như nhau và nằm trong khoảng 10-30 ms (mili giây), hai frame liên tiếp chồng lên nhau với độ dài như nhau và nằm trong khoảng 5-15 ms.

Trên mỗi khung thời gian ngắn frame_t, ta tính các hệ số tương quan sau:

$$\forall k = \overline{0, M-1}, R_{k,t} \overset{def}{=} \sum_{i=0}^{M-1-k} frame_t(i)\, frame_t(k+i) \tag{1}$$

ở đây $\{\text{frame}_t(i)\}_{i=0}^{M-1}$ là các mẫu tín hiệu tiếng nói của frame_t.

Mệnh đề 1. $\forall k = \overline{1, M-1}, R_{k,t} \le R_{0,t}$

Chứng minh:

$$R_{k,t} = \sum_{i=0}^{M-1-k} frame_t(i)\, frame_t(k+i) \le \sum_{i=0}^{M-1-k} \frac{\{frame_t(i)\}^2 + \{frame_t(k+i)\}^2}{2} =$$

$$\frac{1}{2}\sum_{i=0}^{M-1-k}\{frame_t(i)\}^2 + \frac{1}{2}\sum_{i=0}^{M-1-k}\{frame_t(k+i)\}^2 \le \frac{1}{2}\sum_{i=0}^{M-1}\{frame_t(i)\}^2 + \frac{1}{2}\sum_{i'=0}^{M-1}\{frame_t(i')\}^2 = \frac{R_{0,t}}{2} + \frac{R_{0,t}}{2} = R_{0,t}$$

Vậy $R_{k,t} \le R_{0,t}$.

Do frame_t nằm trên đoạn tiếng nói hữu thanh, $\{\text{frame}_t(i)\}_{i=0}^{M-1}$ là tựa tuần hoàn tại chỉ số k_0 tức là $\text{frame}_t(i) \approx \text{frame}_t(i+k_0)$ $\forall i = \overline{0, M-1-k_0}$, khi đó $R_{k_0,t} \approx R_{0,t}$.

Dựa trên mệnh đề 1 ta có thể ước lượng F0 tại khung tính hiệu ngắn frame_t như sau: $F_0(t) \approx \frac{f_s}{k_{0,t}}$; với t là chỉ số dãy frame, f_s là tần số lấy mẫu của tín hiệu tiếng nói (thường $fs \in [8000, 22050]$) và $k_{0,t} \overset{\text{def}}{=} \arg\max_{1 \le k \le M-1}\{R_{k,t}\}$.

Thuật toán trên tuy đơn giản, nhưng độ chính xác phụ thuộc vào tham số độ dài của một frame (với tiếng nói có tần số F0 thấp như giọng nam trầm chẳng hạn, tham số độ dài cửa sổ phải đủ lớn), ngoài ra khi có nhiều giá trị k mà $R_{k,t} \approx R_{k_0,t}$ và $k_{0,t}$ không là đỉnh đúng độ chính xác của thuật toán trên không cao [13].

Thuật toán RAPT [14] ước lượng F0 sử dụng các hệ số tự tương quan được chuẩn hóa,

$$NCCF_t(k) = \frac{R_k}{\sqrt{e_{0,t} e_{k,t}}} \text{ , trong đó } e_{k,t} = \sum_{i=0}^{M-1-k} frame_t(k+i)^2 \quad (2)$$

sau đó tìm đỉnh (chỉ số k, tại đó giá trị tín hiệu đạt cực đại địa phương) của dãy số $\{NCCF_t(k)\}$ và sử dụng thuật toán quy hoạch động để chọn đỉnh phù hợp.

Nói chung một thuật toán ước lượng F0 đều cần biến đổi tín hiệu tiếng nói ban đầu (trong miền thời gian hoặc miền tần số) để làm nổi bật các đỉnh của tín hiệu, sau đó lựa chọn đỉnh hợp lý từ các đỉnh đã tìm được.

Dựa trên giá trị F0 của từng frame trên đoạn hữu thanh, chúng ta có thể xác định các điểm PM như sau :

$pm(t) = pm(t-1) + f_s / f_0(t)$ — $frame_t$ thuộc phần hữu thanh, fs là tần số lấy mẫu, f0(t) là tần số F0 của $frame_t$.

$pm(t) = pm(t-1) + \Delta$ — $frame_t$ thuộc phần vô thanh, Δ là tham số độ rộng cửa sổ. (3)

Tất nhiên với âm thanh đích chưa xác định với đường cách điệu $f_0(t)_{t=1}^{T}$ mong muốn thì chỉ có thể xác định các PM của âm thanh đích bằng cách trên, trái lại khi âm thanh nguồn đã biết, chúng ta có thể xác định các PM trực tiếp từ sóng âm, xem chẳng hạn thuật toán của Talkin được cài đặt trong bộ công cụ [ESPS, Praat]. Chú ý rằng thuật toán của Talkin tính các PM cũng khá phức tạp, tương tự phương pháp tính F0 đã đề cập ở trên.

B. Mô hình Xu

Mô hình Xu [7-9] đã được sử dụng rộng rãi cho tiếng Trung Quốc phổ thông –Mandarin để mô hình hóa đường tần số cơ bản F0 của các thanh điệu trong ngữ cảnh. Từ một số nhỏ tham số của mô hình, chúng ta tạo sinh đường F0 mới có đường nét thích ứng cho các hiện tượng biến thanh của câu sẽ được phát âm. Các tác giả đã đề ra khái niệm "pitch target" trong tiếp cận mô hình hóa đường F0, trong thực tế ứng dụng thường "pitch target" là dạng tuyến tính, và về thực chất mô hình đã biểu diễn sai số xấp xỉ tuyến tính của các giá trị F0 bằng một xấp xỉ mới có dạng là hàm phân rã dạng hàm mũ có tốc độ suy giảm là một giá trị nằm trong khoảng (0,1) [7, 8].

Nói cách khác, các giá trị F0 của tín hiệu khi phát âm âm tiết/từ,… được xấp xỉ bằng hàm có dạng:

$$F(t) \approx \alpha e^{-\lambda t} + at + b. \quad (4)$$

Như vậy F0 được tạo ra từ sự kết hợp của thành phần xấp xỉ tuyến tính at + b và thành phần phân rã hàm mũ $\alpha e^{-\lambda t}$ của sai số xấp xỉ tuyến tính. Trong biểu diễn trên a là độ dốc của đường tuyến tính đích, khi a > 0, đường tuyến tính đích tăng và ngược lại khi a < 0 thì đường tuyến tính đích giảm, λ>0 và $\alpha e^{-\lambda t}$ biểu thị tốc độ phân rã dạng hàm mũ của sai số xấp xỉ tuyến tính của đường F0.

Nếu gán sai số [7],

$$y(t) \overset{\text{def}}{=} F(t) - (at+b), \quad y(t+1) \approx \alpha e^{-\lambda(t+1)} = \underset{k \in (0,1)}{e^{-\lambda}} \alpha e^{-\lambda t} \approx k y(t). \quad (5)$$

Như vậy khi khớp (fitting) từng tập dữ liệu F0, chúng ta có thể rút gọn việc tìm các tham số hồi quy phi tuyến a, b, α và λ khá phức tạp bằng việc tính các tham số hồi quy a,b và k, trong đó k∈ (0,1).

III. ĐỀ XUẤT

Trong phần này chúng tôi đề xuất 2 thuật toán tương ứng với hai bước để tổng hợp thanh điệu của âm tiết tiếng Việt phát âm rời.

A. Thuật toán ước lượng các PM dựa trên tổng tích lũy

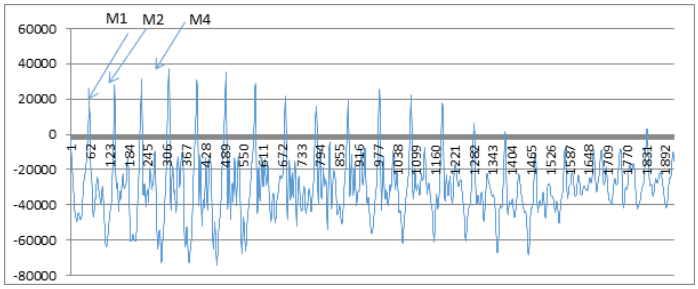
Trên đoạn tiếng nói (trong miền thời gian hoặc miền tần số), $x = \{x_j\}_{1 \le j \le N}$ chúng ta xác định dãy tổng tích lũy $s = \{s_j\}_{1 \le j \le N}$, trong đó

$$s_1 = x_1, \ \forall j = \overline{2, N}, s_j \overset{\text{def}}{=} s_{j-1} + x_j \quad (6)$$

và chúng ta cũng xác định chỉ số không điểm-giao (ZC, zero crossing) như sau:

$$Z_x(m) = \frac{1}{M} \sum_{n=m-M+1}^{m} \frac{|sign(x_n) - sign(x_{n-1})|}{2}, m = \overline{M+1, N} \quad (7)$$

ở đây M là tham số mẫu đứng trước và $sign(x)=\begin{cases}1, x>0\\0, x=0\\-1, x<0\end{cases}$



Hình 2. Dãy tổng tích lũy của một phát âm cho âm tiết "s/á/u"

Các giá trị $Z_x(m)$ dùng để xác định mẫu tiếng nói thuộc đoạn âm thanh hoặc hữu thanh, khi $Z_x(m)>\alpha_z$ thì mẫu x_m thuộc phần vô thanh và ngược lại, tham số α_z được xác định bằng thực nghiệm (chẳng hạn, có thể chọn α_z=0.17).

Thuật toán được mô tả và thực hiện như sau:

Thuật toán 1. EPM (Ước lượng các giá trị PM của sóng tiếng nói).

Đầu vào: Dãy tín hiệu tiếng nói $\{x_m\}_{1\le m\le N}$ trong miền thời gian.

Tham số α_e, $\alpha_z\in(0,1)$: ngưỡng quyết định vô thanh/hữu thanh
Tham số $f_{0,min}$, $f_{0,max}$ (giá trị nguyên dương, đơn vị Hz)
Tham số tần số lấy mẫu f_s, số mẫu của một frame (đoạn tín hiệu ngắn) là M.

Đầu ra: Số các đoạn hữu thanh K, các điểm pitch mark theo 4 phương pháp lấy

$pm^+_{k,j}\ _{1\le k\le K,1\le j\le n^+_k}$, $pm^-_{k,j}\ _{1\le k\le K,1\le j\le n^-_k}$, $p^-_{k,j}\ _{1\le k\le K,1\le j\le n^-_{k,2}}$, $p_{k,j}\ _{1\le k\le K,1\le j\le n^+_{k,2}}$

Bước 1: Phân đoạn tín hiệu $\{x_m\}_{1\le m\le N}$ thành các đoạn tiếng nói hữu thanh và vô thanh

1.1: Xác định tín hiệu giao không điểm $\{Z_x(m)\}$ theo công thức (7)

1.2: Chia đoạn tín hiệu $\{x_m\}_{1\le m\le N}$ thành các frame độ dài M mẫu, không chồng lên nhau. Xác định năng lượng của từng $frame_t=\{x_m\}_{(t-1)M+1\le m\le tM}$ theo công thức $E_t=\sum_{m=(t-1)M+1}^{tM}x_m^2$

1.3: Loại bỏ nền (silence): Xác định K và các đoạn tín hiệu tiếng nói $\{x_m\}_{M*t_{k,1}\le m\le M*t_{k,2}}, k=\overline{1,K}$ theo ngưỡng năng lượng α_e*max{E_t}, ở đây $t_{k,1}$, $t_{k,2}$ là chỉ số frame nhỏ nhất bắt đầu và lớn nhất kết thúc của mỗi đoạn k, sao cho $E_{t_{k,1}}\ge\alpha_e\max\{E_t\}, E_{t_{k,2}}\ge\alpha_e\max\{E_t\}$.

1.4: Với mỗi đoạn tiếng nói $\{x_m\}_{M*t_{k,1}\le m\le M*t_{k,2}}, k=\overline{1,K}$, trích đoạn tiếng nói hữu thanh $\{x_m\}_{N_{k,1}\le m\le N_{k,2}}$ $x_{N_{k,1}}, x_{N_{k,2}}$ là mẫu x_n bắt đầu và mẫu x_n cuối cùng của đoạn tín hiệu thứ k sao cho $Z_x(n)<\alpha_z$.

Bước 2 : $T_{min}=f_s/f_{0,max}$, $T_{max}=f_s/f_{0,min}$

Bước 3: Lặp trên mỗi đoạn tín hiệu hữu thanh $\{x_m\}_{N_{k,1}\le m\le N_{k,2}}, k=\overline{1,K}$,

3.1: Tính tổng tín hiệu tổng tích lũy $S_k=\{s_m\}_{N_{k,1}\le m\le N_{k,2}}, k=\overline{1,K}$ theo công thức (6)

3.2: Xác định các đỉnh của S_k, tính trung bình biên độ tại các đỉnh của S_k : $mean_k=\dfrac{\sum_{n\in peak\{S_k\}}|S_{k,n}|}{\#\,peak\{S_k\}}$

3.3: Xác định đỉnh phù hợp đầu tiên của S_k, $p_{k,1}$ là đỉnh n đầu tiên mà $S_{k,n}$ trên ngưỡng $mean_k$;

$p_{k,1} = \arg\min_{n \in peak\{S_k\}} |S_{k,n}| \geq mean_k$,

3.4. Loại các đỉnh $n \in peak\{S_k\}, n < p_{k,1}$

3.5: Chỉ giữ lại các đỉnh $p_{k,j} \in peak\{S_k\}, p_{k,j} - p_{k,j-1} \in [T_{k,min}, T_{k,max}]$

3.6: Loại tiếp các đỉnh nhỏ (thứ 2 sát ngay các đỉnh lớn), tức là loại các đỉnh $p_{k,j}$ mà $p_{k,j} < \min p_{k,j-1} \cdot p_{k,j+1}$

3.7: Loại các đỉnh $\frac{fs}{p_{k,j+1} - p_{k,j}} > 0.5 * \left(\frac{fs}{p_{k,j} - p_{k,j-1}} + \frac{fs}{p_{k,j+2} - p_{k,j+1}} \right)$ (Loại các điểm của đường F0 bị nhô lên so với điểm gần nhất bên trái và bên phải).

3.8: Chèn lại một đỉnh

$$m \in peak(S_k), p_{k,j} < m < p_{k,j+1}, \left| p_m - \frac{p_{k,j} + p_{k,j+1}}{2} \right| \to \min \text{ khi } \frac{fs}{p_{k,j+1} - p_{k,j}} < 0.5 * \left(\frac{fs}{p_{k,j} - p_{k,j-1}} + \frac{fs}{p_{k,j+2} - p_{k,j+1}} \right)$$

(Chèn lại một đỉnh đã bị loại tại điểm của đường F0 bị thấp xuống so với điểm gần nhất bên trái và bên phải).

3.9: Đặt $peak_R\{S_k\}$ là tập đỉnh rút gọn.

Bước 4: Lặp, trên mỗi phần hữu thanh $\{x_m\}_{N_{k,1} \leq m \leq N_{k,2}}, k = \overline{1, K}$,

4.1: Chu kỳ F0 trên mỗi đoạn $[p_{k,j}, p_{k,j+1}]$ của đoạn $\{x_m\}_{N_{k,1} \leq m \leq N_{k,2}}$, $f_{0,k,j} = \frac{f_s}{p_{k,j+1} - p_{k,j}}, j = \overline{1, \# peak_R\{S_k\} - 1}$

4.2: Xác định các điểm pitch mark theo tiêu chí điểm cực đại địa phương của đoạn $\{x_m\}_{N_{k,1} \leq m \leq N_{k,2}}$

$$pm^+_{k,j} = max \underset{p_{k,j} \leq n \leq p_{k,j+1}}{peak} x_n$$

4.3: Xác định các điểm pitch mark theo tiêu chí điểm cực tiểu địa phương của đoạn $\{x_m\}_{N_{k,1} \leq m \leq N_{k,2}}$

$$pm^-_{k,j} = \min \underset{p_{k,j} \leq n \leq p_{k,j+1}}{peak} -x_n$$

Bước 5: Xác định các điểm Pulse kiểu Praat[Praat, tương tự bước 3 nhưng lấy điểm cực tiểu địa phương của tín hiệu tổng tích lũy, ta được $p^-_{k,j}$ trên mỗi phần hữu thanh $\{x_m\}_{N_{k,1} \leq m \leq N_{k,2}}, k = \overline{1, K}$.

Bước 6: Kết thúc

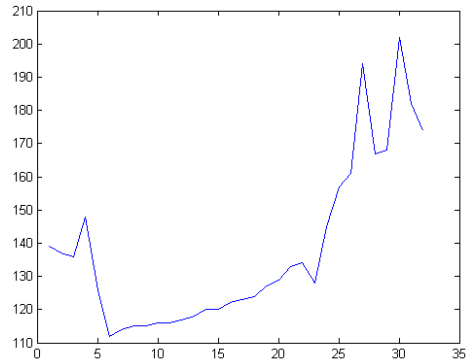
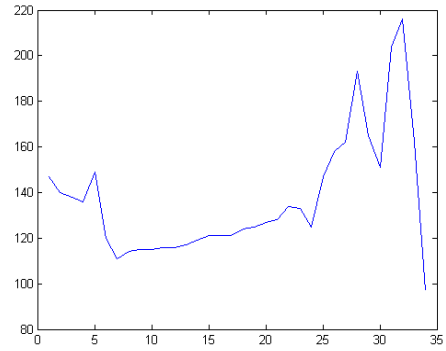
Trả về: Số các đoạn hữu thanh K, các điểm pitch mark theo 4 phương pháp lấy

$pm^+_{k,j}{}_{1 \leq k \leq K, 1 \leq j \leq n^+_k}$, $pm^-_{k,j}{}_{1 \leq k \leq K, 1 \leq j \leq n^-_k}$, $p^-_{k,j}{}_{1 \leq k \leq K, 1 \leq j \leq n^-_{k,2}}$, $p_{k,j}{}_{1 \leq k \leq K, 1 \leq j \leq n^+_{k,2}}$

Rõ ràng trong thuật toán chúng ta chỉ sử dụng các phép tỉnh tổng và các phép so sánh nên có thể thấy độ phức tạp thuật toán 1 là nhỏ. M là độ dài của một khung tiếng nói ngắn nên M<=N (N là số mẫu tín hiệu tiếng nói đầu vào), vì vậy độ phức tạp của thuật toán 1 là O(N).

Thuật toán duyệt tuần tự từng khung tiếng nói (khi tính các giá trị năng lượng) và duyệt từng mẫu tín hiệu tiếng nói khi tính các giá trị đổi dấu của tín hiệu trên từng đoạn tiếng nói trích rút được nên Thuật toán luôn dừng do số lượng mẫu tín hiệu đầu vào đã được xác định (Thuật toán không làm việc với câu tiếng nói thu nhận trong thời gian thực)

Nhận xét: Từ các PM tìm được chúng ta sẽ có một ước lượng thô của các giá trị F0, $f_{0,k,j}$ theo công thức ở bước 4.1 thuật toán 1, tuy nhiên để có một ước lượng tốt của giá trị đường F0 còn phải hậu xử lý rất nhiều, chẳng hạn phép làm trơn các giá trị đường F0.

(a)

Hình 3. *Đường F0 thô, không trơn* $f_s/(p_{j+1}-p_j)$ *của âm /bốn/* , (a) {p_j} *là các điểm Pulse tính bằng Praat* [20], và (b) *là các điểm giao cắt 0 chuyển từ âm sang dương tính bằng thuật toán 1 đề xuất.*

B. Thuật toán sử dụng phương pháp bình phương tối thiểu để tính các tham số hồi quy phi tuyến mô hình Xu

Việc tính các hệ số của mô hình Xu khi cho trước giá trị đường F0 cũng sử dụng phương pháp bình phương tối thiểu, thay vì tìm các hệ số a, b, α, λ ta xác định các hệ số a,b, k ∈ (0,1) (k=$e^{-\lambda}$) bằng phép cực tiểu hóa như sau:

$$F(a,b,k)=\sum_{i=1}^{n-1}\left[f_{0,i+1}-a(i+1)-b-k\left(f_{0,i}-ai-b\right)\right]^2\rightarrow\min \tag{8}$$

trong đó n là số frame của đoạn tiếng nói, $f_{0,i}{}_{i=1}^{n}$ là giá trị đường F0 của đoạn tiếng nói.

Áp dụng phương pháp bình phương tối thiểu, tham số a, b và k cần tìm thỏa mãn hệ phương trình:

$$\begin{cases}\frac{\partial F}{\partial a}=0\\ \frac{\partial F}{\partial b}=0\\ \frac{\partial F}{\partial k}=0\end{cases}\Rightarrow\begin{cases}\sum_{i=0}^{n-1} i\left[f_{0,i+1}-a(i+1)-b-k(f_{0,i}-ai-b)\right]=0\\ \sum_{i=0}^{n-1}\left[f_{0,i+1}-a(i+1)-b-k(f_{0,i}-ai-b)\right]=0\\ \sum_{i=0}^{n-1} f_{0,i}\left[f_{0,i+1}-a(i+1)-b-k(f_{0,i}-ai-b)\right]=0\end{cases}\Rightarrow\begin{cases}\sum_{i=0}^{n-1} if_{0,i+1}-a\sum_{i=0}^{n-1} i(i+1)-b(1-k)\sum_{i=0}^{n-1} i-k\sum_{i=0}^{n-1} if_{0,i}+ak\sum_{i=0}^{n-1} i^2=0\\ \sum_{i=0}^{n-1} f_{0,i+1}-a\sum_{i=0}^{n-1}(i+1)-k\sum_{i=0}^{n-1} f_{0,i}+ak\sum_{i=0}^{n-1} i-nb(1-k)=0\\ \sum_{i=0}^{n-1} f_{0,i}f_{0,i+1}-a\sum_{i=0}^{n-1}(i+1)f_{0,i}-k\sum_{i=0}^{n-1} f_{0,i}^2+ak\sum_{i=0}^{n-1} if_{0,i}-b(1-k)\sum_{i=0}^{n-1} f_{0,i}=0\end{cases}$$

Đặt

$$\begin{aligned}&S_0\overset{def}{=}\sum_{i=0}^{n-1} f_{0,i}\overset{def}{=}n\overline{f}_0\Rightarrow S_1\overset{def}{=}\sum_{i=0}^{n-1} f_{0,i+1}=S_0+f_{0,n}-f_{0,0};S_2\overset{def}{=}\sum_{i=0}^{n-1} if_{0,i}\Rightarrow S_3\overset{def}{=}\sum_{i=0}^{n-1}(i+1)f_{0,i+1}=S_2+nf_{0,n};\\ &\sum_{i=0}^{n-1} if_{0,i+1}=S_3-S_1;\sum_{i=0}^{n-1}(i+1)f_{0,i}=S_2+S_0\end{aligned} \tag{9}$$

$$S_{yy}\overset{def}{=}\sum_{i=0}^{n-1} f_{0,i}f_{0,i+1},S_{y,2}\overset{def}{=}\sum_{i=0}^{n-1} f_{0,i}^2 \tag{10}$$

Do $\forall l\in N,\sum_{i=0}^{l} i=\frac{l(l+1)}{2},\sum_{i=0}^{l} i^2=\frac{l(l+1)(2l+1)}{6}$ ta có

$$6(S_3-S_1)-a(n-1)n(2n+2)-3(n-1)nb(1-k)-6kS_2+ak(n-1)n(2n-1)=0 \tag{11}$$

$$2S_1-an(n+1)-2kS_0+akn(n-1)=2nb(1-k) \tag{12}$$

$$S_{yy}-a(S_2+S_0)-kS_{y,2}+akS_2-b(1-k)n\overline{F}_0=0 \tag{13}$$

Kết hợp (11) và (12), (13) và (12) ta được a(1-k) và k thỏa mãn hệ phương trình sau:

$$-a(1-k)(n-1)n(n+1)+k\left[-12S_2+6(n-1)S_0\right]=6(n+1)S_1-12S_3 \tag{14}$$

$$-a(1-k)\left[2S_2-S_0(n-1)\right]+k\left[-2S_{yy}+2S_0\overline{F}_0\right]=2F_0S_1-2S_{yy} \tag{15}$$

Thuật toán 2. XNF, tính tham số hồi quy phi tuyến cho mô hình Xu, với phương trình hồi quy dạng (8).

Đầu vào: Dãy giá trị F0 $\{f_{0,i}\}_{0\le i\le n}$ của đoạn sóng âm.

Đầu ra: Số thực a,b và k, trong đó 0 <k< 1 sao cho hàm $F_{a,b,k}$ $\{f_{0,i}\}_{0\le i\le n}$ đạt giá trị nhỏ nhất

Bước 1: Tính S_0, $\overline{f}_0$, S_1, S_2, S_3, S_{yy}, $S_{y,2}$ theo công thức (9).

Bước 2: Tính a(1-k), k theo công thức (14) và (15), từ đó suy ra a,k.

Bước 3: Tính b theo công thức (12)

Trả về: a, b và k

Do một âm tiết tiếng Việt thường dài không quá 500 mili giây và do đó thường n không vượt quá 30 nếu xét một frame dài khoảng 20 mili giây, nên chúng ta bỏ qua không ước lượng độ phức tạp của thuật toán 2.

Sau khi xác định được các hệ số hồi quy phi tuyến a, b và k đường F0 được sinh bằng công thức:

$$d \overset{def}{=} f_{0,0}^{old} - b,\ f_{0,t}^{new} = a*t + b + k^{t-1}d, t = \overline{1,n}, \text{ ở đây (n+1) là số frame.} \quad (16)$$

IV. THỰC NGHIỆM

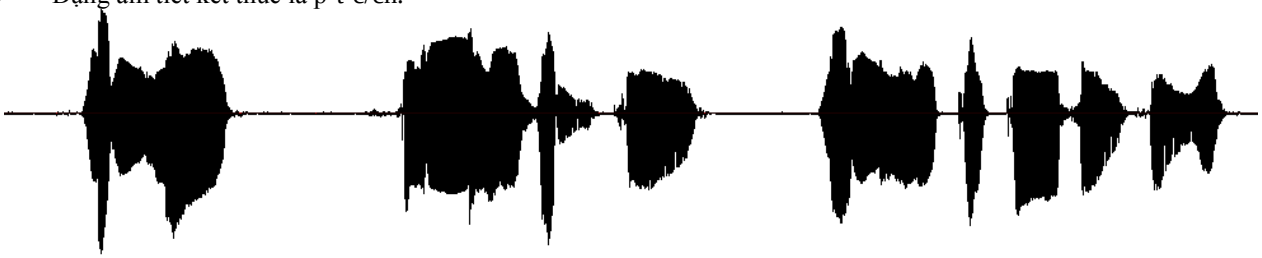
Thực nghiệm được tổ chức để thể hiện các ưu điểm của thuật toán 1 và sự hiệu quả của phép giải của hồi quy phi tuyến trong vấn đề tổng hợp thanh điệu cho âm tiết rời thuật toán 2.

A. Dữ liệu thực nghiệm

Dữ liệu thực nghiệm gồm phát âm các âm tiết mang thanh điệu điển hình của một giọng nam để phân tích hiệu quả của cách ước lượng các điểm PM, một tập các phát âm của âm tiết mang các thanh tiếng Việt để cách điệu hóa đường F0 điển hình của các thanh theo thuật toán 2. Tiếng nói được thu nhận trong môi trường văn phòng độ ồn thấp, tần số lấy mẫu là fs = 11025 Hz, định dạng 16bit, đơn kênh.

Các âm tiết được chọn để phát âm gồm hơn 70 âm tiết rời, giọng nam và 581 giọng nữ (phát thanh viên chuyên nghiệp), câu văn bản được chọn từ truyện dế mèn phiêu lưu ký bao gồm đầy đủ các dạng đặc tính ngữ âm của âm tiết tiếng Việt:

- Dạng âm tiết chỉ có nguyên âm
- Dạng âm tiết không có phụ âm đầu
- Dạng âm tiết kết thúc là bán nguyên âm.
- Dạng âm tiết kết thúc là p-t-c/ch.

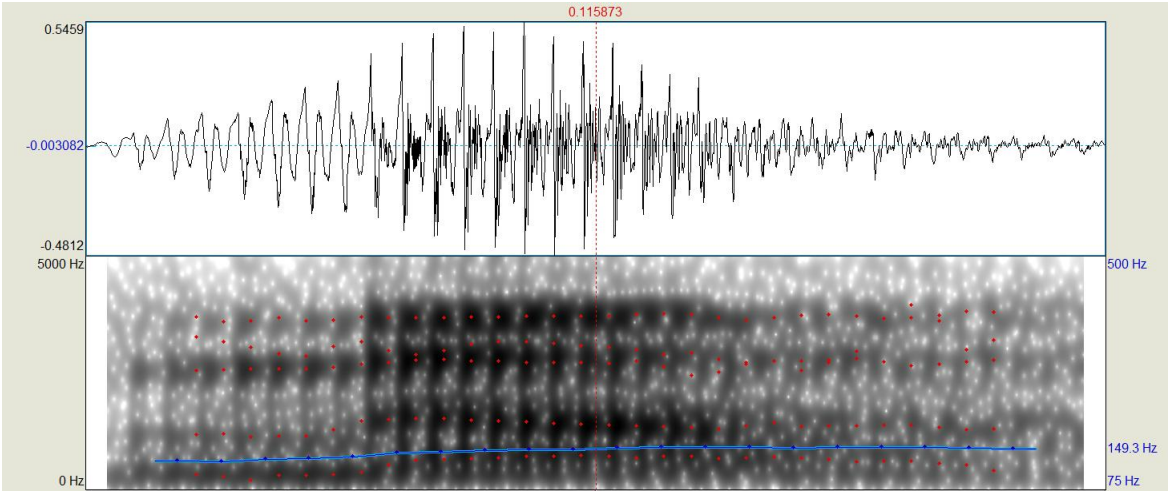


Hình 4. Câu "Đừng lo, xem mây vận trời đêm nay có cơ đổi gió." – Trích truyện dế mèn Phiêu lưu ký

B. Thử nghiệm xác định các điểm PM

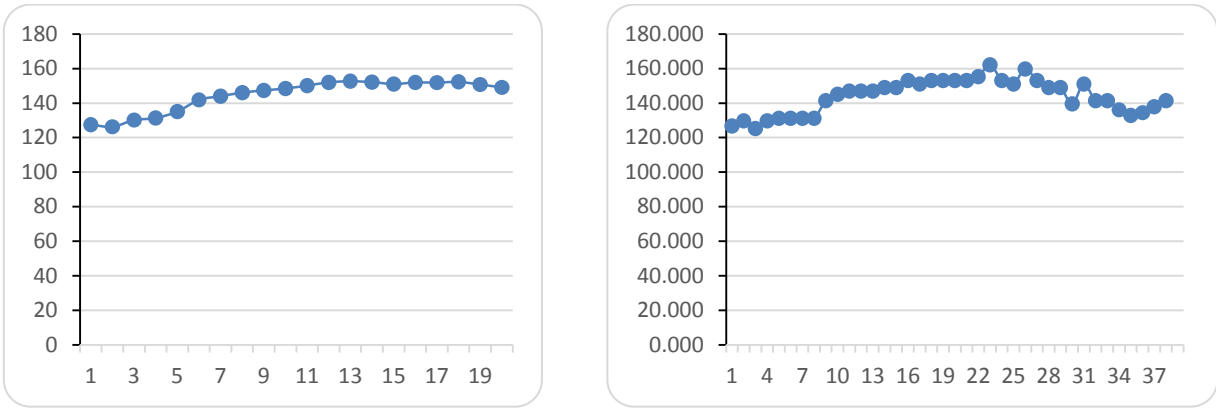
Các công thức cho thấy thuật toán 1 có khối lượng tính toán nhỏ hơn so với thuật toán kiểu dùng quy hoạch động [13] để tối ưu hóa cách lựa chọn các PM từ tập các đỉnh của sóng tiếng nói. Về độ tin cậy của thuật toán 1, chúng tôi sẽ so sánh thuật toán 1 với thuật toán kiểu Talkin được cài đặt trong phần mềm Praat [20].

Tham số $f0_{min}$ = 50 Hz, $f0_{max}$ = 550 Hz và ngưỡng loại bỏ âm nền α_e = 0.2 và ngưỡng quyết định hữu thanh/vô thanh α_z = 0.17 (để loại bỏ các phần tín hiệu vô thanh thuộc phụ âm đầu và âm cuối).



Hình 5. Xác định các Pulse âm /năm/ giọng nam bằng Praat [20]

So với phương pháp của thuật toán 1 sử dụng các điểm cực đại địa phương của tổng tích luỹ các giá trị mẫu âm thanh từ đó đưa ra giá trị chu kỳ-tần số, Praat thực hiện việc tính giá trị tần số trong các khoảng thời gian 10ms.



a-*Dãy giá trị F0 từ Praat* b-*Dãy giá trị F0 từ Thuật toán 1*

Hình 6. So sánh các dãy giá trị F0 từ Praat và b-Thuật toán 1

Đối sánh song song các kết quả, chúng ta thấy một số vấn đề như sau:

- Praat tính giá trị tần số trong các khoảng thời gian 10 ms nên các khoảng thời gian tính đều nhau, số liệu tần số đưa ra được làm trơn hơn.
- Tuy nhiên, trên thực tế Praat để mất nhiều dữ liệu khi phân tích âm thanh. Với tần số âm thanh 140 Hz – ta có tương đương khoảng thời gian một chu kỳ $t_0 = 1000/140 \approx 7{,}143$ ms. Việc tính toán trong các khoảng thời gian 10 ms đã làm mất khoảng 1/3 dữ liệu tín hiệu thực tế, chưa kể phần bị loại bỏ trong quá trình sửa lỗi. Do đó, trong ví dụ so sánh với Thuật toán 1 - xác định trực tiếp các điểm PM và chu kỳ, tần số tương ứng – ta thấy số lượng F0 của Praat thu được ít hơn so với Thuật toán 1.
- Cách tính theo thuật toán sử dụng trong Thuật toán 1 chỉ ra cụ thể các điểm PM, chỉ rõ được các chu kỳ tín hiệu âm thanh. Trên cơ sở đó có thể phân tích kỹ hơn các đặc tính có thể có của âm thanh, phục vụ tốt hơn cho mục đích *Phân tích – Tổng hợp* âm thanh tiếng nói.

C. Tổng hợp đường F0 điển hình của các thanh điệu tiếng Việt theo mô hình Xu với phép giải hồi quy phi tuyến của thuật toán 2.

Các kết quả tổng hợp thanh điệu được thực nghiệm cho các âm tiết được thu nhận để bao gồm các dạng âm tiết tiếng Việt như:

- Chỉ có nguyên âm
- Không có phụ âm đầu
- Kết thúc là bán nguyên âm.
- Kết thúc là p-t-c/ch,...

Để tổng hợp đường F0 điển hình của một thanh điệu trước khi tổng hợp một âm tiết phát âm rời mang thanh tương ứng từ một âm tiết mang thanh ngang (với các âm tiết không tận cùng là p-t-c/ch) hoặc thanh nặng (với các âm

tiết tận cùng là p-t-c/ch), chúng tôi đã xây dựng một công cụ biên tập đường F0 và giải hồi quy phi tuyến của thuật toán 2 đề xuất Các bước tiến hành phân tích bao gồm:

Bước 1. Tính đường nét F0 của âm tiết mang thanh điệu

Bước 2. Biên tập lại các giá trị F0 bằng cách kéo nhả, lưu ý tại các vị trí điển hình của đường F0

+ *Giá trị F0 ban đầu*
+ *Giá trị F0 kết thúc*
+ *Giá trị F0 bắt đầu đi lên*
+ *Giá trị F0 bắt đầu đi xuống*

Bước 3. Tính các hệ số hồi quy a, b và k của đường F0 theo thuật toán 2.

Bước 4. Tổng hợp đường nét thanh điệu mới sinh được bằng công thức (16).

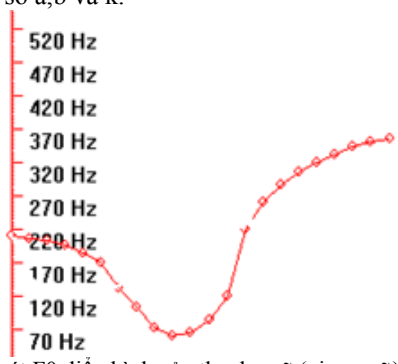
Bước 5. Xác định các điểm PM của âm tiết mang đường nét F0 mới sẽ được tổng hợp ra dựa trên công thức (3)

Bước 6. Xác định các PM từ sóng âm tiết của âm tiết đầu vào, mang thanh ngang (với các âm tiết không tận cùng là p-t-c/ch) hoặc thanh nặng (với các âm tiết tận cùng là p-t-c/ch) theo thuật toán 1.

Bước 7. Sử dụng phương pháp PSOLA để chuyển sóng âm tiết đầu vào thành sóng âm tiết có đường nét F0 sinh ra ở bước 4.

Bước 8. Cảm nhận bằng tai câu nói tổng hợp, so sánh với câu nói gốc và điều chỉnh lại.

Bước 9. Lưu lại các giá trị F0 bắt đầu, hệ số a,b và k.



Hình 7. Đường nét F0 điển hình của thanh ngã (giọng nữ) sau khi biên tập

Một lưu ý về đặc điểm đường nét của đường F0 điển hình của các thanh điệu tiếng Việt khi phát âm các âm tiết rời (xem bảng 2) sẽ cho chúng ta sử dụng một bộ hệ số hồi quy {a, b, k} cho các thanh {nganh, huyền, sắc, nặng} và dùng 2 bộ tham số {a, b, k} cho thanh ngã và thanh hỏi.

Bảng 2. Hình dáng thanh điệu tiếng Việt phát âm rời [19]

Thanh	Âm vực	Âm điệu	Hướng	Đường nét F0 điển hình
Không dấu	Cao	Bằng	Không đổi	Đi ngang
Huyền	Thấp	Bằng	Không đổi	Thanh huyền được phát âm ở âm vực thấp hơn so với thanh ngang. Đường nét đi xuống thoai thoải đều từ đầu đến phần cuối âm tiết.
Sắc	Cao	Trắc	Không đổi	Đường nét đi lên phụ thuộc vào loại hình âm tiết và độ dài ngắn của nguyên âm. Đường F0 đi ngang đến giữa vần và sau đó đi lên ở phần cuối âm tiết.
Nặng	Thấp	Trắc	Không đổi	Đường F0 đi xuống đột ngột ở phần cuối âm tiết
Hỏi	Thấp	Trắc	Đổi hướng	Đường F0 bắt đầu với mức cao của thanh huyền, đi xuống thoai thoải đến giữa vần, đi ngang một thời gian ngắn và sau đó lại đi lên cân đối với đường đi xuống. Giá trị F0 ở đầu và cuối âm tiết xấp xỉ bằng nhau.
Ngã	Cao	Trắc	Đổi hướng	Đường F0 bắt đầu ở độ cao cao hơn thanh huyền, thấp hơn thanh ngang. Ở giữa âm tiết giảm xuống đột ngột, hoặc có hiện tượng bị đứt, sau đó lại tăng lên đạt tần số cao nhất ở cuối âm tiết.
Sắc (âm tận cùng p-t-c/ch)	Thấp	Bằng	Không đổi	Đi ngang
Nặng (âm tận cùng p-t-c/ch)	Thấp	Bằng	Không đổi	Đi ngang

Các âm tiết đã được biến đổi thanh được các chuyên gia ngữ âm kiểm tra cho thấy là âm nghe rõ, không bị hiện tượng rè, thanh điệu tổng hợp nghe rõ ràng, giữ được đường nét đặc trưng thanh điệu tương ứng, đặc biệt phương pháp đề xuất đã tổng hợp tốt các thanh ngã và thanh nặng.

Phương pháp cách điệu đường F0 đề xuất không cần đi vào chi tiết xây dựng luật như phép cách điệu tuyến tính trong [2], điều này là do phép hồi quy tuyến tính tạo ra sai số lớn rất khó để tạo ra đường cách điệu bám sát phần đi xuống cũng như phần đi lên của đường F0 của sóng âm tiết mang các thanh ngã và thanh nặng. Ngoài ra, phương pháp đề xuất có thể mở rộng sang việc tổng hợp đường F0 của sóng âm từ tiếng Việt (từ đa âm tiết). Trong các nghiên cứu tiếp theo chúng tôi sẽ mở rộng thuật toán 2 để khảo sát hiện tượng biến thanh trên từ đa âm tiết của tiếng Việt.

V. KẾT LUẬN

Bài báo đã đề xuất hai thuật toán trong vấn đề tổng hợp tiếng nói với các đóng góp mới như sau:

- Thứ nhất - thuật toán xác định các điểm đánh dấu pitch của tín hiệu tiếng nói gốc trong miền thời gian dựa trên tín hiệu tổng tích lũy. Thuật toán đơn giản, không cần phép chia thành các đoạn ngắn (frame) như các phương pháp khác, đạt độ chính xác cao với tín hiệu âm tiết đầu vào dùng cho tổng hợp thanh điệu Việt. Với dữ liệu âm thanh tiếng Việt của các âm tiết và ngữ đoạn được thử nghiệm đã bao hàm đầy đủ hiện tượng ngữ âm tiếng Việt, kết quả tính các điểm đánh dấu pitch theo tiếp cận mới đã chứng tỏ sự đúng đắn, chính xác tương đương và hơn tiếp cận tính kiểu Praat.
- Thứ hai - thuật toán xác định các tham số hồi quy phi tuyến theo tiếp cận pitch target của mô hình Xu để sinh đường nét F0 điển hình của các thanh điệu tiếng Việt. Thuật toán hồi quy phi tuyến đơn giản, kết quả sinh đường F0 điển hình của các thanh điệu là khá tốt (kể cả các âm tiết tận cùng p-t-c/ch), chất lượng tổng hợp thanh điệu cũng vẫn tốt với các thanh ngã, thanh nặng của các âm tiết tiếng Việt phát âm rời.

Trong các nghiên cứu tiếp theo chúng tôi sẽ mở rộng kết quả để tổng hợp thanh điệu cho các từ đa âm tiết trong tiếng Việt với các điểm đánh dấu pitch được lựa chọn tự động từ tiếp cận tổng tích lũy của chúng tôi.

TÀI LIỆU THAM KHẢO

[1] Trịnh Anh Tuấn, Nghiên cứu các đặc trưng để phân tích và tổng hợp tín hiệu âm tần, Luận án tiến sỹ, Học viện Công nghệ Bưu chính Viễn thông.

[2] Tu Trong Do, Tomio Takara, Vietnamese Tones Generation Using F0 and Power Patterns, ICASSP 2003. DOI: 10.1109/ICASSP.2003.1198828.

[3] Hansjoerg Mixdorf, Nguyen Tien Dung, Lương Chi Mai (2003), Quantitative Analysis and Synthesis of Syllabic Tones in Vietnamese, Proc. in EUROSPEECH, 177-180.

[4] Dung, Mixdorff, et al (2004), Fujisaki Model based F0 contours in Vietnamese TTS, Proceedings of ICSLP2004.

[5] Hansjoerg Mixdorf, Nguyen Tien Dung, Lương Chi Mai, Ngo Hoang Huy, Vu Kim Bang (2004), Toward integrating the Fujisaki model into Vietnamese TTS, Proceeding of the International Conference on Spoken Language Processing, Korea.

[6] Nghiên cứu cải tiến đặc tính thanh điệu của hệ thống tổng hợp tiếng Việt, Phan Thanh Sơn, Dương Tử Cường, Vũ Tất Thắng, Lương Chi Mai. Fair 2013.

[7] Ching X. Xu*, Yi Xu*, and Li-Shi Luo, A PITCH TARGET APPROXIMATION MODEL FOR F0 CONTOURS IN MANDARIN, ICPhS99.

[8] Xu, Y., and Wang, Q. E. in press. Pitch targets and their realization: Evidence from Mandarin. Speech Communication 33 (2001) 319-337.

[9] Yi Xu, Tonal alignment, syllable structure and coarticulation: Toward an integrated model, Italian Journal of Linguistics (2006) 18: 125-159.

[10] Yi Xu1, Santitham Prom-on, Articulatory-Functional Modeling of Speech Prosody: A Review CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION 2010 (INTERSPEECH 2010), VOLS 1-4. (pp. 46 - 49). ISCA-INST SPEECH COMMUNICATION ASSOC.

[11] Xu, Y. and Prom-on, S. (2014). Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. Speech Communication 57, 181-208.

[12] Santitham Prom-on and Yi Xu, DISCOVERING UNDERLYING TONAL REPRESENTATIONS BY COMPUTATIONAL MODELING: A CASE STUDY OF THAI.Phonology 32 2015: 505-535. DOI: https://doi.org/10.1017/S0952675715000299.

[13] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)" in "Speech Coding & Synthesis", W B Kleijn, K K Paliwal eds, Elsevier ISBN 0444821694, 1995.

[14] Hynek Bořil and Petr Pollák , DIRECT TIME DOMAIN FUNDAMENTAL FREQUENCY ESTIMATION OF SPEECH IN NOISY CONDITIONS, 1003-1006, Eusipco 2004.

[15] Dongmei Wang, John H.L. Hansen , F0 ESTIMATION FOR NOISY SPEECH BY EXPLORING TEMPORAL HARMONIC STRUCTURES IN LOCAL TIME FREQUENCY SPECTRUM SEGMENT ICASSP 2016.

[16] A Short Guide to Pitch-marking in the Festival Speech Synthesis System and Recommendations for Improvements.

[17] M. Leg´at, J. Matouˇsek, D. Tihelka. On the Detection of Pitch Marks Using a Robust MultiPhase Algorithm. Speech Communication, Elsevier : North-Holland, 2011, 53 (4), pp.552.

[18] Đoàn Thiện Thuật (2000), Ngữ âm tiếng Việt, NXB Đại học Quốc gia Hà Nội.

[19] http://www.speech.kth.se/wavesurfer/links.html

[20] http://www.praat.org

AN EFFICIENT METHOD FOR DETERMINING PITCH MARKS AND NONLINEAR REGRESSION FOR VIETNAMESE TONAL SYNTHESIS

Ta Yen Thai, Nguyen Van Hung, Ngo Hoang Huy, Pham Kim Thu

***ABSTRACT**: In the study of Vietnamese synthesis, synthesis of tones plays an important role and shows the uniqueness of Vietnamese. Mixed-tone algorithms in Vietnamese often use the least squares method to simulate the linearity of F0 lines of Vietnamese tones that are discrete or continuous in the dialect. Access in this direction is very difficult to stylize the linear F0 of some Vietnamese tones such as heavy bar and bar.*

Besides, the effective estimation of the pitch mark (PM) of a pronunciation is the first important step in the synthesis of tones so far that is still an open question. This paper presents a method for estimating the PM of speech waves and the method of constructing a nonlinear regression for the synthesis of tones based on the Xu model which has been widely used for Chinese - Mandarin. Experimentation has proved the effectiveness of the algorithm proposed for synthesizing Vietnamese tone tone discrete. The synthesized tones sounded with both the heavy bar and the fallen bar, keeping the corresponding tone characteristic. In addition, every syllable in Vietnamese (vowel, no consonant, ending in vowel or end is p-t-c / ch) are synthesized well hear, no phenomenon fade…

***Keywords**: Voiced/unvoiced decision, Zerocrossing, Pitch mark F0 contour stylization, Xu model, Tone synthesis.*