Tổng hợp tiếng Việt có cảm xúc

Lê Xuân Thành[1], Trịnh Văn Loan[1], Nguyễn Hồng Quang[1], Đào Thị Lệ Thủy[1,2], Đinh Đồng Lưỡng[3]
[1]Viện Công nghệ Thông tin và Truyền thông, Trường Đại học Bách khoa Hà Nội
[2]Khoa Công nghệ Thông tin, Trường Cao đẳng nghề Công nghệ cao Hà Nội
[3]Khoa Công nghệ Thông tin, Trường Đại học Nha Trang
E-mail: thanhlx@soict.hust.edu.vn, loantv@soict.hust.edu.vn, quangnh@soict.hust.edu.vn, thuydt@hht.edu.vn, quangnh@soict.hust.edu.vn
Tác giả liên hệ: Lê Xuân Thành
Ngày nhận: 06/11/2017, ngày sửa chữa: 11/12/2017, ngày duyệt đăng: 28/12/2017

Tóm tắt: Tiếng Việt là ngôn ngữ đơn âm tiết và có thanh điệu. Để tổng hợp tiếng Việt chất lượng tốt, việc đảm bảo chất lượng của thanh điệu tổng hợp sao cho càng gần với thanh điệu tự nhiên là rất quan trọng. Bài báo này đề xuất một phương pháp tổng hợp tiếng Việt dựa trên ghép nối âm vị kép, trong đó các biến thiên $F0$ của các âm được tổng hợp giống như biến thiên $F0$ của tiếng nói tự nhiên. Hơn nữa, để tích hợp cảm xúc vào tiếng Việt tổng hợp, bài báo trình bày một phương pháp tổng hợp dựa trên mô hình Fujisaki. Ba cảm xúc khác nhau được thử nghiệm là buồn, tức và vui. Các kết quả đánh giá khách quan và chủ quan chất lượng tiếng Việt tổng hợp cũng được trình bày trong nghiên cứu này.

Từ khóa: *Tiếng Việt, tổng hợp, thanh điệu, cảm xúc, ghép nối, Fujisaki.*

Title: **Synthesis of Emotional Vietnamese**

Abstract: Vietnamese is a monosyllabic and tonal language. To synthesize good quality Vietnamese, the quality of synthesized tones, which is ideally close to that of natural speech, is very important. This paper proposes a concatenation-based synthesis method for Vietnamese in which the variations of $F0$ of the synthesized tones are as similar as natural voice. Furthermore, in order to integrate emotions into the synthesized speech, the paper presents a synthesis method based on Fujisaki model. Three different emotions are investigated, including sadness, anger, and happiness. Objective and subjective evaluations are presented in this study.

Keywords: *Vietnamese, synthesis, tone, emotion, concatenation, Fujisaki.*

I. GIỚI THIỆU

Tổng hợp tiếng nói nói chung [1, 2] và tổng hợp tiếng nói có cảm xúc nói riêng [3, 4], đã được nghiên cứu từ lâu trong các ngôn ngữ khác như tiếng Anh [5], tiếng Đức [6], tiếng Hà Lan [7], tiếng Thụy Điển [8], v.v.

Trong tiếng Việt, nghiên cứu về tổng hợp tiếng nói đã có nhiều kết quả tốt. Có thể kể đến các nghiên cứu của nhóm của Lương Chi Mai, nghiên cứu ảnh hưởng của $F0$ đến thanh điệu [9, 10] bằng mô hình Fujisaki, tổng hợp theo phương pháp mô phỏng tham số bằng mô hình Markov ẩn (HMM: Hidden Markov Model) [11]; hay các nghiên cứu đến từ Viện MICA, Trường Đại học Bách khoa Hà Nội về tổng hợp theo phương pháp ghép nối [12], ảnh hưởng của $F0$ đến tiếng nói tổng hợp [13], tổng hợp sử dụng mô hình HMM [14].

Các nghiên cứu về tổng hợp tiếng Việt có cảm xúc chưa nhiều. Các nghiên cứu này đều có một số kết quả bước đầu nhưng cũng tồn tại một số vấn đề sau đây: không thuần túy phân tích ảnh hưởng của các tham số như $F0$, thời hạn, năng lượng đến cảm xúc song lại kết hợp giữa tiếng nói và các dữ liệu hình ảnh, hoặc là các kết quả nghiên cứu thực hiện trên các bộ ngữ liệu còn hạn chế về số lượng cũng như chưa đi sâu vào nghiên cứu các cảm xúc cơ bản.

Có thể kể đến một vài kết quả nghiên cứu kết hợp giữa tiếng nói và các dữ liệu video, hình ảnh biểu hiện khuôn mặt, cử chỉ, các tín hiệu điện não, v.v.

Các nghiên cứu trong [15, 16] thử nghiệm tổng hợp tiếng Việt có cảm xúc bằng các mô hình hóa ngôn điệu tiếng Việt với ngữ liệu đa thể thức. Nhóm nghiên cứu Thi Duyen Ngo [17] đã sử dụng ngữ liệu có cảm xúc bao gồm các phát âm tiếng Việt của một nam nghệ sỹ và một nữ nghệ sỹ, phát âm 19 câu ở năm cảm xúc: tự nhiên, vui, buồn, hơi giận, rất giận. Một số tác giả Trung Quốc như LaVutuan [18], Jiang [19] đã kết hợp với sinh viên Việt Nam xây dựng ngữ liệu cảm xúc tiếng Việt theo cách đóng kịch biểu lộ sáu cảm xúc: vui, bình thường, buồn, ngạc nhiên, tức, sợ hãi, kết hợp với dữ liệu cảm xúc tiếng Trung Quốc nhằm nghiên cứu chéo các tham số ảnh hưởng đến cảm xúc trong hai ngôn ngữ.

Bảng I
CÁCH TỔ CHỨC ĐƠN VỊ ÂM ĐẦU VÀ ĐƠN VỊ ÂM CUỐI

Đơn vị âm đầu		**Đơn vị âm cuối**		
Thanh ngang		Đầy đủ 6 thanh điệu		
Âm đầu	Âm đệm	Âm đệm	Âm chính	Âm cuối

Để góp phần nghiên cứu cảm xúc của tiếng Việt nói, bài báo này trình bày một số giải pháp như sau. Trước hết, chúng tôi đề xuất mô hình tổng hợp tiếng Việt chất lượng tốt để tổng hợp được các câu nói với cảm xúc bình thường và mục tiêu cao nhất là giữ được chất lượng thanh điệu tự nhiên để phục vụ cho tổng hợp tiếng nói có cảm xúc. Tiếp theo, chúng tôi sử dụng kết quả xây dựng bộ ngữ liệu cảm xúc tiếng Việt (BKEmo [20]) để xây dựng mô hình tổng hợp tiếng Việt có cảm xúc bằng cách điều chỉnh các tham số thời hạn, cường độ với công cụ Praat [21], và điều chỉnh quy luật biến thiên $F0$ theo mô hình Fujisaki. Cuối cùng, tiếng Việt tổng hợp được đánh giá chủ quan bằng sử dụng người nghe trực tiếp và khách quan bằng so sánh phổ. Đối với phương pháp đánh giá chủ quan, người nghe tham gia đánh giá là các sinh viên đại học đã được học môn Xử lý tiếng nói của ngành Công nghệ Thông tin nên đã có kiến thức về tiếng nói tổng hợp và phương pháp chủ quan đánh giá chất lượng tiếng nói. Kết quả đánh giá cho thấy, hệ thống tổng hợp tiếng Việt khá tốt ở cảm xúc bình thường, buồn và tức, và sau đó là cảm xúc vui.

Mục II của bài báo sẽ trình bày những nội dung cơ bản của việc xây dựng bộ ngữ liệu tiếng Việt và xây dựng bộ tổng hợp tiếng Việt có chất lượng tốt. Mục III trình bày khái quát việc xây dựng ngữ liệu tiếng Việt có cảm xúc, chi tiết các đề xuất, thuật giải để tổng hợp tiếng Việt có cảm xúc, và kết quả đánh giá chất lượng tiếng Việt có cảm xúc đã được tổng hợp. Cuối cùng, mục IV là kết luận.

II. TỔNG HỢP TIẾNG VIỆT CHẤT LƯỢNG TỐT

1. Xây dựng ngữ liệu cho bộ tổng hợp tiếng Việt chất lượng tốt

Phần này của bài báo trình bày kết quả xây dựng bộ tổng hợp tiếng Việt với mục tiêu chất lượng thanh điệu là quan trọng nhất, chiếm vị trí hàng đầu để phục vụ tổng hợp tiếng Việt nói có cảm xúc.

Phương pháp tổng hợp bằng ghép nối âm vị kép đã được sử dụng. Đầu tiên, xây dựng ngữ liệu là bước rất quan trọng trong quá trình tạo nên bộ tổng hợp tiếng Việt chất lượng tốt. Phương án xây dựng bộ ngữ liệu mới của tiếng Việt được đề nghị như sau: một âm tiết bất kỳ trong tiếng Việt được chia thành âm đầu và âm cuối (Bảng I). Trong đó, thời hạn của âm đầu sẽ được xác định từ điểm bắt đầu của âm tiết tới phần ổn định của nguyên âm trong âm tiết đó. Âm cuối được xác định từ điểm bắt đầu ổn định của nguyên âm trong âm tiết đến hết âm tiết. Cách làm này đảm bảo mỗi âm tiết chỉ cần xử lý một điểm ghép nối duy nhất tại vùng ổn định của nguyên âm có trong âm tiết. Ví dụ âm tiết "*bàng*" sẽ được chia thành: phần âm đầu */ba/* và âm cuối */àng/*. Để đảm bảo tính tự nhiên của thanh điệu, các thanh điệu sẽ được giữ nguyên như đã được ghi âm và thuộc về âm cuối. Âm đầu sẽ chỉ chứa thanh ngang còn âm cuối sẽ chứa đầy đủ cả 6 thanh điệu (Bảng I). Ví dụ: âm đầu */ta/* kết với các âm cuối */án/*, */àn/*, */an/*, */ản/*, */ãn/*, */ạn/* để tạo nên các âm tiết "*tán*", "*tàn*", "*tan*", "*tản*", "*tãn*", "*tạn*". Từ đó, cần tính toán để xây dựng kịch bản thu phù hợp đảm bảo ngữ liệu đầy đủ thỏa mãn yêu cầu đề ra và chọn giọng để thu, tổ chức kịch bản thu để có chất lượng tốt nhất.

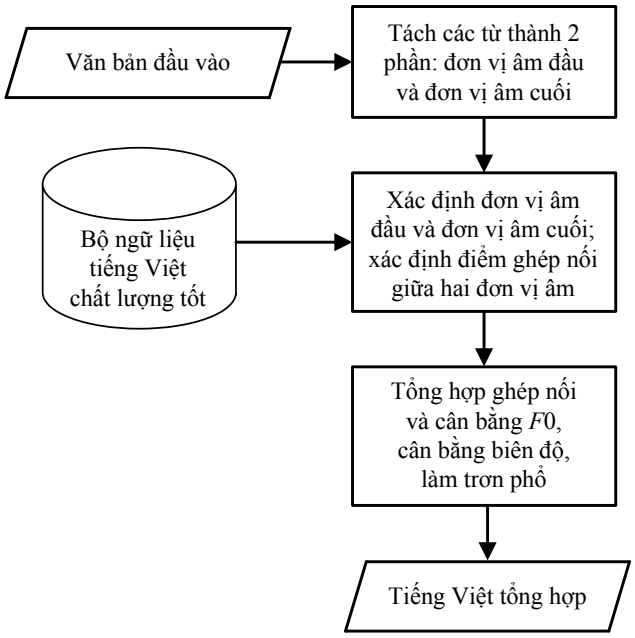
Bước đầu, tiến hành ghi âm cho bốn giọng: một giọng nam, một giọng nữ và hai giọng trẻ em. Tín hiệu thu được lấy mẫu ở tần số 16000 Hz và 16 bit cho một mẫu. Thời gian thu mỗi bộ 1015 âm tiết liên tục là 50,75 phút (tính cả khoảng lặng giữa các âm tiết). Tổng dung lượng của 1015 âm tiết là 98 MB cho mỗi giọng. Đây là bộ ngữ liệu xây dựng để phục vụ cho mục đích nghiên cứu. Với các ứng dụng thực tế, nếu tách lấy đơn vị âm đầu và đơn vị âm cuối dùng cho tổng hợp và phần còn lại được cắt bỏ thì dung lượng sẽ giảm đi. Theo kết quả tính toán, tỷ số tín hiệu trên nhiễu trung bình của bộ ngữ liệu đã được xây dựng là 38 dB. Đây là kết quả tốt chấp nhận được.

2. Tổng hợp tiếng Việt chất lượng tốt bằng phương pháp ghép nối

Các phương pháp tổng hợp tiếng nói hiện nay cơ bản được chia thành hai hướng: tổng hợp tiếng nói trực tiếp và tổng hợp tiếng nói dựa trên mô hình [22, 23], trong đó tổng hợp tiếng nói trực tiếp thường cho chất lượng cao vì bản thân tiếng nói tự nhiên đã được dùng trực tiếp để tổng hợp. Trong nghiên cứu này, phương pháp tổng hợp trực tiếp dựa trên các đơn vị âm đầu và đơn vị âm cuối được chọn từ tiếng nói ghi âm. Đây là phương pháp cho chất lượng tiếng nói tổng hợp khá tự nhiên, đặc biệt là chất lượng thanh điệu vì các thanh điệu được giữ nguyên như tiếng nói tự nhiên.

1) Tổng hợp bằng phương pháp ghép nối:

Quá trình tổng hợp tiếng Việt bằng phương pháp ghép nối được trình bày trên Hình 1. Theo quá trình này, để tổng hợp một âm tiết, đầu tiên cần xác định âm đầu và âm cuối để ghép nối. Điểm ghép nối cần được chọn thuộc vùng ổn định của nguyên âm thuộc âm sẽ tổng hợp. Các âm đầu và âm cuối của bộ tổng hợp đã được lựa chọn trong quá trình xây dựng bộ ngữ liệu. Vì vậy, trong bộ ngữ liệu đã có sẵn các âm này cùng với vị trí của điểm ghép nối. Bộ tổng hợp thực hiện ghép nối các âm và thực hiện các thuật giải cân bằng và làm trơn tham số tại điểm ghép nối.



Hình 1. Lưu đồ bộ tổng hợp tiếng Việt bằng phương pháp ghép nối.

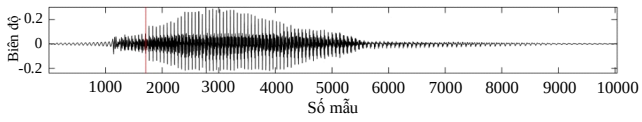
2) Cân bằng tham số tại vị trí ghép nối:

Quá trình ghép nối âm đầu với âm cuối, thực hiện các bước cân bằng biên độ, làm trơn $F0$ và phổ tại điểm ghép nối được thể hiện thông qua Hình 1. Văn bản đầu vào sẽ được tách từ và gán nhãn theo quy luật được trình bày ở phần xây dựng bộ ngữ liệu. Âm đầu và âm cuối được lựa chọn trong bộ ngữ liệu. Bộ tổng hợp tiến hành ghép nối hai âm này, thực hiện cân bằng biên độ, cân bằng $F0$ và làm trơn phổ tại điểm ghép nối.

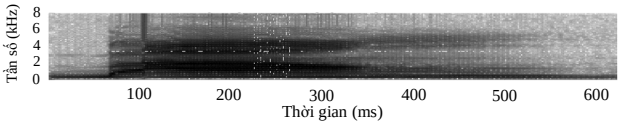
Quá trình cân bằng và làm trơn các tham số được minh họa cho trường hợp tổng hợp âm "*bàng*" như sau. Âm cần tổng hợp "*bàng*" sẽ được ghép âm đầu trích từ tập tin chứa âm /*ba*/ có âm cuối trích từ tập tin chứa âm "*àng*". Âm đầu /*ba*/ có tần số cơ bản $F0_1 = 266,67$ Hz. Tần số $F0$ của âm cuối /*àng*/ sau khi tách là $F0_2 = 213,33$ Hz.

Nếu ghép nối một cách đơn giản mà không có thao tác cân bằng biên độ, làm trơn $F0$ và phổ tại điểm ghép nối, sẽ có dạng sóng tín hiệu tổng hợp "*bàng*", spectrogram và biến thiên $F0$ như Hình 2. Có thể thấy sự chênh lệch biên độ của âm đầu và âm cuối tại điểm ghép nối (Hình 2(a)) và biến thiên gãy khúc của $F0$ theo thời gian (Hình 2(c)).

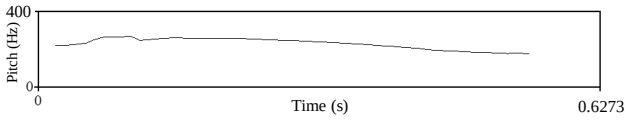
Việc cân bằng tại điểm ghép nối được thực hiện theo thuật giải TD-PSOLA [24] trong đó $F0$ của âm đầu cần được giảm xuống để cân bằng với $F0$ của đoạn âm cuối. Biên độ tín hiệu của đoạn âm đầu cần được tăng lên để biên độ biến thiên trơn tại vùng ghép nối. Sau khi thực hiện cân bằng $F0$ và biên độ, tín hiệu âm "*bàng*" được trình bày trên Hình 3. Có thể thấy biến thiên biên độ và biến thiên $F0$ đã không còn đột biến ở điểm ghép nối.



(a) Dạng sóng của âm "*bàng*" được tổng hợp bằng cách ghép đơn giản

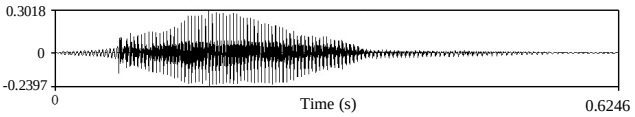


(b) Spectrogram của âm "*bàng*" sau khi tổng hợp đơn giản

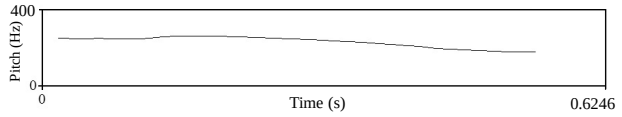


(c) Biến thiên $F0$ của âm "*bàng*" đã được tổng hợp đơn giản

Hình 2. Tín hiệu của âm "*bàng*" khi chưa xử lý điểm ghép nối



(a) Dạng sóng âm "*bàng*" sau khi cân bằng biên độ



(b) Biến thiên $F0$ theo thời gian sau khi cân bằng $F0$

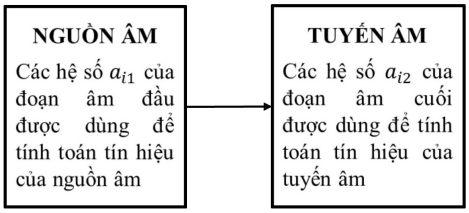
Hình 3. Tín hiệu của âm tiết "*bàng*" sau khi cân bằng biên độ và cân bằng $F0$.

Sau khi cân bằng biên độ và tần số cơ bản, để cải thiện tiếng nói tổng hợp, cần làm trơn phổ tại vùng ghép nối. Mã hóa tiên đoán tuyến tính (LPC: Linear Prediction Coding) [25] đã được sử dụng để làm trơn phổ tại vùng ghép nối. Bài báo đề xuất phương pháp làm trơn như sau: Tín hiệu nguồn âm của đoạn âm đầu sẽ kích thích cho tuyến âm của đoạn âm cuối ở vị trí ghép nối để tạo ra tín hiệu âm đầu mới. Tín hiệu của nguồn âm của đoạn âm đầu và tham số tuyến âm của đoạn âm cuối ở vị trí ghép nối được xác định bằng LPC như mô tả trên Hình 4. Cụ thể các bước như sau:

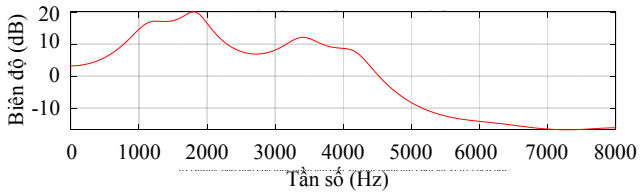
Tham số a_{i1}, $i = 1, \ldots, P$, $P = 12$, sẽ được sử dụng để tính tín hiệu nguồn âm để kích thích cho tuyến âm bằng công thức

$$e(n) = y(n) + \sum_{i=1}^{P} a_{i1} y(n-i), \tag{1}$$

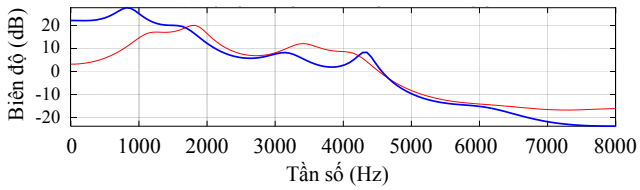
trong đó $y(n)$ là tín hiệu tiếng nói của âm đầu.



Hình 4. Sơ đồ khối quá trình làm trơn phổ.



(a) Đường bao phổ đoạn âm cuối tại điểm ghép nối



(b) Đường bao phổ của đoạn âm cuối (nét mảnh) và đường bao phổ của đoạn âm đầu (nét đậm) tại vị trí ghép nối

Hình 5. Đường bao phổ của âm tiết "*bàng*" trước khi được cân bằng phổ.

Tín hiệu tổng hợp $y_1(n)$ được tổng hợp dựa trên công thức

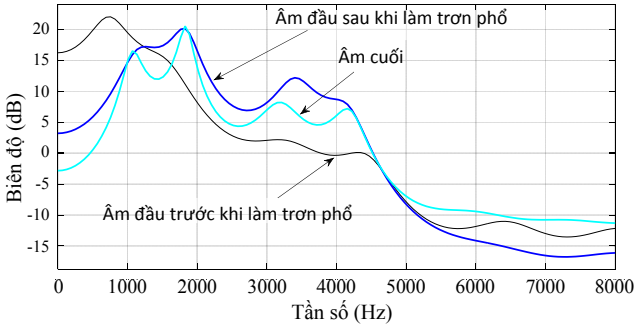
$$y_1(n) = e(n) - \sum_{i=1}^{P} a_{i2} y(n-i), \quad (2)$$

trong đó tín hiệu kích thích chính là $e(n)$ trong công thức (1) và các tham số của tuyến âm a_{i2}, $i = 1, \ldots, P$, là của phần âm cuối. Tín hiệu $y_1(n)$ chính là tín hiệu tiếng nói của âm đầu đã được cân bằng phổ.

Hình 5 biểu diễn đường bao phổ âm đầu trước khi làm trơn phổ. Hình 5(a) là đường bao phổ của một phần âm cuối tại vị trí ghép nối và được vẽ trên Hình 5(b) cùng với đường bao phổ của đoạn âm đầu để so sánh. Hình 5(b) cho thấy chênh lệch khá lớn giữa hai đường bao phổ này trước khi tiến hành làm trơn phổ.

Từ Hình 6 có thể thấy, việc làm trơn phổ của vùng ghép nối nói chung đã giảm đi nhiều chênh lệch đường bao phổ của đoạn âm cuối so với đoạn âm đầu.

Đo lường khoảng cách phổ (trình bày ở mục III-5) của đoạn âm cuối với đoạn âm đầu tại vị trí ghép nối trong trường hợp này cho thấy: trước khi làm trơn thì khoảng cách này trung bình là 4,7%, sau khi làm trơn trung bình khoảng cách giảm xuống còn 2,89%.



Hình 6. Đường bao phổ của âm đầu và một phần âm cuối tại điểm ghép nối trước và sau khi làm trơn bằng LPC.

Bảng II
CÁC CÂU ĐƯỢC TỔNG HỢP

TT	Nội dung
1	Cảnh vật chung quanh tôi đều thay đổi
2	Nhìn chúng tôi với cặp mắt hiền từ và cảm động
3	Cũng may, đã có tiếng dạ rang của phụ huynh đáp lại
4	Một cậu đứng đầu ôm mặt khóc
5	Một mùi hương lạ xông lên trong lớp
6	Để thầy, mẹ được vui lòng, các em phải cố gắng học
7	Các em đã nghe chưa
8	Mấy cậu học trò lớp ba cũng đua nhau quay đầu nhìn ra
9	Không thể nào quên được những cảm giác trong sáng ấy
10	Một buổi mai đầy sương thu và gió lạnh

3. Đánh giá kết quả chất lượng tiếng Việt tổng hợp ở mức câu

Phương pháp đánh giá chủ quan dùng điểm trung bình số ý kiến (MOS: Mean Opinion Score) [26] đã được lựa chọn để đánh giá chất lượng tiếng Việt tổng hợp bằng phương pháp ghép nối của nghiên cứu này.

Để phục vụ cho bộ tổng hợp tiếng Việt có cảm xúc sẽ trình bày ở mục III, chất lượng của các câu nói ở giọng trần thuật (cảm xúc bình thường) được quan tâm. Trong thử nghiệm này, 10 câu nói có nội dung được liệt kê trong Bảng II đã được tổng hợp và đánh giá.

Người nghe được yêu cầu nghe từng câu tổng hợp được phát ngẫu nhiên sau đó đánh giá theo thang điểm 5 của thang MOS với các điểm từ 1 đến 5 lần lượt là: rất kém, kém, bình thường, tốt và rất tốt.

Bảng III là kết quả đánh giá do 14 sinh viên của cùng một lớp thực hiện. Kết quả đánh giá các câu đều ở mức tốt, trong đó câu 4 được đánh giá với điểm số cao nhất, câu 3 và câu 8 có kết quả thấp do là câu khá dài nên việc điều chỉnh các tham số chưa tốt lắm.

Bảng III
ĐIỂM ĐÁNH GIÁ CỦA 14 NGƯỜI NGHE

Câu	Điểm TB cộng	Câu	Điểm TB cộng
Câu 1	3,1429	Câu 6	3,2143
Câu 2	2,8571	Câu 7	3,1429
Câu 3	2,5000	Câu 8	2,7143
Câu 4	4,1429	Câu 9	3,6429
Câu 5	3,7857	Câu 10	3,2143

III. TỔNG HỢP TIẾNG VIỆT CÓ CẢM XÚC

1. Xây dựng ngữ liệu cho bộ tổng hợp tiếng Việt có cảm xúc

Bộ ngữ liệu cảm xúc tiếng Việt BKEmo được xây dựng bước đầu cho 4 cảm xúc cơ bản: vui, buồn, tức, bình thường. Kịch bản thu được xây dựng để các diễn viên chuyên nghiệp thể hiện được 4 cảm xúc một cách tự nhiên nhất theo cùng một cách biểu cảm đã được thảo luận trước. Kịch bản thu cho bộ ngữ liệu BKEmo được xây dựng với trợ giúp của các nhà ngôn ngữ của Viện Ngôn ngữ Việt Nam.

Bộ ngữ liệu BKEmo gồm 56 giọng được thu âm (28 nữ và 28 nam), có độ tuổi trải đều từ 18 đến 60 tuổi, các diễn viên có kinh nghiệm biểu đạt khá tốt, rõ ràng cảm xúc khi thu. Các cảm xúc được biểu diễn theo một cách thống nhất (cùng một kiểu vui, cùng một kiểu buồn, v.v.), dễ nhận ra hay dễ biểu lộ nhất để đảm bảo số lượng dữ liệu đủ lớn giúp tìm ra quy luật.

Bộ ngữ liệu BKEmo được thu trong phòng thu âm, lồng tiếng chuyên nghiệp có hệ thống cách âm, lọc nhiễu tốt. Mỗi câu được lưu thành một tệp tin wav, tín hiệu thu được lấy mẫu ở tần số 16000 Hz và 16 bit cho một mẫu. Mỗi người nói sẽ có 220 tệp tin cho một cảm xúc. Ngữ liệu thu được gồm có 52800 tệp tin với tổng dung lượng là 2,68 Gb.

Bộ ngữ liệu này đã được đánh giá [20] bằng phương pháp nghe trực tiếp và phân tích các đặc trưng của cảm xúc trong tiếng Việt nói [20, 27, 28]. Các tham số đặc trưng này đã được sử dụng để nhận dạng nhằm đánh giá chất lượng của bộ ngữ liệu [20, 27, 28] trước khi được dùng cho tổng hợp tiếng Việt có biểu lộ cảm xúc. Ở đây, chủ yếu sẽ thay đổi 3 tham số cơ bản là $F0$, cao độ của giọng nói và tốc độ phát âm để có các cảm xúc khác nhau.

Kết quả của [20] cho thấy tần số cơ bản $F0$ trung bình cho cảm xúc buồn là thấp nhất, tiếp theo là cảm xúc bình thường. Cảm xúc tức và cảm xúc vui có $F0$ lớn hơn so với cảm xúc buồn và cảm xúc bình thường. Cảm xúc tức có giá trị $F0$ trung bình lớn nhất. Về mặt năng lượng, cảm xúc buồn và cảm xúc tức có độ chênh lệch năng lượng lớn nhất. Với giọng nữ, không có chênh lệch nhiều về năng lượng giữa cảm xúc bình thường và cảm xúc buồn, còn chênh lệch giữa cảm xúc vui và buồn có độ chênh lệch năng lượng cao nhất.

Thông thường, với cảm xúc tức, tốc độ phát âm có thể nhanh hơn so với cảm xúc bình thường hay cảm xúc buồn. Cảm xúc buồn có cao độ thấp hơn hẳn so với cảm xúc tức nhưng khá gần với cảm xúc bình thường, tốc độ nói của cảm xúc vui khá gần với cảm xúc tức.

2. Mô hình Fujisaki trong tổng hợp tiếng nói

Theo kết quả nghiên cứu trong [20, 27, 28], tần số cơ bản $F0$ đóng vai trò rất quan trọng trong việc thể hiện các cảm xúc. Một vài tham số khác như cường độ và thời hạn sẽ kết hợp với $F0$ góp phần nâng cao chất lượng thể hiện cảm xúc. Tiếng Việt là ngôn ngữ có thanh điệu nên mô hình Fujisakiđã được lựa chọn để điều chỉnh $F0$ khi tổng hợp tiếng Việt có cảm xúc. Mô hình Fujisaki đã được thử nghiệm với nhiều ngôn ngữ khác nhau như tiếng Nhật [29], tiếng Thổ Nhĩ Kỳ [30], tiếng Tây Ban Nha [31], tiếng Bồ Đào Nha [32], tiếng Trung Quốc [33], v.v. với ưu điểm là sự đơn giản và can thiệp mạnh mẽ vào các tham số tổng hợp tiếng nói [34, 35], hỗ trợ rất tốt cho tổng hợp bằng phương pháp ghép nối [36]. Hình 7 là mô hình Fujisaki dùng cho tổng hợp tiếng Việt có thanh điệu [37].

3. Tổng hợp tiếng Việt có cảm xúc bằng phương pháp ghép nối sử dụng mô hình Fujisaki

Nội dung tiếp theo của bài báo sẽ trình bày phương pháp tổng hợp tiếng Việt có cảm xúc dựa trên việc thay đổi các tham số cường độ và tốc độ phát âm bằng công cụ Praat, và thay đổi $F0$ bằng mô hình Fujisaki.

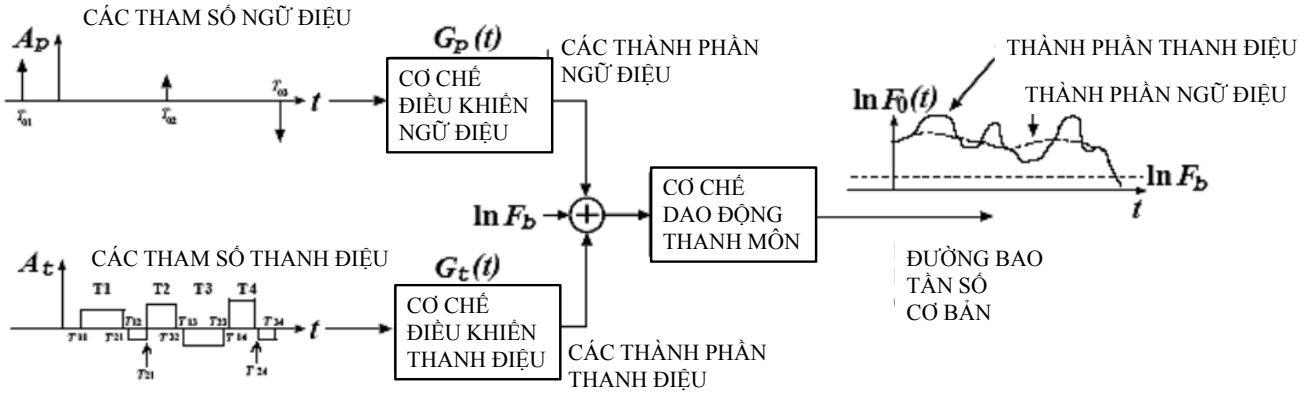
Trong mô hình ở Hình 7, tiếng nói nói chung trên thế giới khi được cảm thụ sẽ có 2 khái niệm: ngữ điệu và trọng âm. Ngữ điệu là cho cả câu (phrase), còn trọng âm (accent) thường cho một âm tiết (có thể một từ trong câu). Tương ứng với tiếng Việt có ngữ điệu và thanh điệu (tone). Biến thiên $F0$ theo thời gian sẽ xác định ngữ điệu và trọng âm (thanh điệu) của câu nói. Mô hình này mô tả quy luật biến thiên tần số cơ bản $F0$ theo thời gian cho tiếng Việt có thanh điệu.

Trong mô hình trên, tập các tần số $F0$ sẽ được sinh ra khi điều chỉnh các tham số của 3 công thức dưới đây:

$$\ln F0_t = \ln F_b + \sum_{j=1}^{I} Ap_i Gp(t - T_{oi}) + \sum_{k=0}^{n} Aa_i \left[Ga(t - T_{1j}) - Ga(t - T_{2j}) \right], \quad (3)$$

$$Gp(t) = \begin{cases} \alpha^2 t e^{-\alpha t}, & t \geq 0, \\ 0, & t < 0, \end{cases} \quad (4)$$

$$Ga(t) = \begin{cases} \min\left[1 - (1 + \beta t)e^{-\beta t}, \gamma\right], & t \geq 0, \\ 0, & t < 0. \end{cases} \quad (5)$$

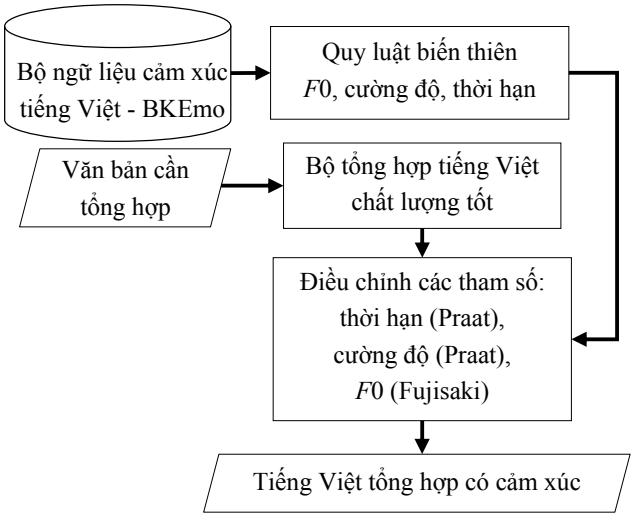


Hình 7. Mô hình Fujisaki áp dụng cho tổng hợp tiếng Việt có thanh điệu [37].

Ba công thức trên bao gồm các thành phần sau:

- **Thành phần 1**: F_b là hằng số, thường xác định theo giới tính hoặc lứa tuổi, chẳng hạn, giọng nam có F_b thấp hơn giọng nữ, trẻ em có F_b cao hơn người lớn; α là hằng số đóng vai trò là tần số góc tự nhiên của lệnh ngữ (phrase command); β là tần số góc tự nhiên của lệnh trọng âm (accent command); γ là hằng số chỉ mức giá trị trần tương ứng với các thành phần trọng âm; các đối số bao gồm: I là số lệnh ngữ, J là số lệnh trọng âm; T_0 là thời điểm bắt đầu lệnh ngữ; T_1 và T_2 là thời điểm bắt đầu và kết thúc thanh điệu ở lệnh trọng âm.
- **Thành phần 2**: Bao gồm Ap_i và Gp_i, đây là thành phần quyết định biến thiên $F0$ cho cả câu (phrase command). Về mặt công thức toán, đây được xem là công thức tính tín hiệu ra của một bộ lọc trong đó tín hiệu vào là các hệ số Ap còn đáp ứng xung là Gp. Tín hiệu vào chỉ là các xung Delta (xung đơn vị) như thấy trong hình vẽ mô hình. Vậy đầu ra của thành phần này cho biết quy luật biến thiên $F0$ của câu.
- **Thành phần 3**: Bao gồm Aa và Ga, đây là thành phần cho biết quy luật biến thiên của trọng âm với các ngôn ngữ không có thanh điệu, còn với tiếng Việt chính là quy luật biến thiên $F0$ của thanh điệu. Đây cũng là công thức tính tín hiệu ra của một bộ lọc trong đó đáp ứng xung coi là Ga, còn các Aa là tín hiệu vào (chú ý: về mặt xử lý tín hiệu có thể đổi lẫn tín hiệu vào và đáp ứng xung cho nhau, nên gọi một trong hai đại lượng này, đại lượng nào là tín hiệu vào cũng được, khi đó đại lượng còn lại là đáp ứng xung). Vì vậy, ta nói đấy là thành phần ứng với tín hiệu bậc đơn vị và trên hình vẽ là các tín hiệu vào có dạng chữ nhật dương hoặc âm (với tiếng Việt).

Cộng xếp chồng cả 3 thành phần trên cho quy luật biến thiên $F0$ của cả câu nói bao gồm ngữ điệu và thanh điệu đối với tiếng Việt. Cách xác định các tham số của 3 thành phần này cho tiếng Việt như sau: giọng nam có F_b = 124 Hz, giọng nữ có F_b = 215 Hz, còn lại các thành phần khác được xác định theo thuật giải được công bố trong kết quả nghiên cứu của [10, 37]. Trong nghiên cứu này, α và β được thiết lập tương ứng là 2 Hz và 25 Hz, còn γ được lựa chọn bằng 0,9.



Hình 8. Lưu đồ thuật giải tổng hợp tiếng Việt có cảm xúc bằng phương pháp ghép nối.

Quá trình tổng hợp tiếng Việt có biểu lộ cảm xúc được thực hiện theo quy trình được mô tả trên Hình 8. Văn bản đầu vào sẽ được gắn nhãn cảm xúc, tách từ tự động và được tổng hợp thành câu nói với cảm xúc bình thường. Sau đó câu tổng hợp được điều chỉnh các tham số: thời hạn phát âm và cường độ bằng công cụ Praat, điều chỉnh $F0$ bằng mô hình Fujisaki theo quy luật đã được công bố trong [20, 27, 28].

Bảng IV
MA TRẬN NHẦM LẪN GIỌNG NỮ VỚI CÂU "ÔNG NÓI GÌ THẾ TÔI KHÔNG HIỂU"

	Bình thường	**Tức**	**Vui**	**Buồn**
Bình thường	24	0	0	1
Tức	4	18	3	0
Vui	13	6	5	1
Buồn	8	0	1	16

Bảng V
MA TRẬN NHẦM LẪN GIỌNG NAM VỚI CÂU "ÔNG NÓI GÌ THẾ TÔI KHÔNG HIỂU"

	Bình thường	**Tức**	**Vui**	**Buồn**
Bình thường	20	1	0	4
Tức	1	23	0	1
Vui	7	12	6	0
Buồn	12	1	0	12

Có 600 câu được tổng hợp với 4 cảm xúc: vui, buồn, bình thường và tức. Trong đó: cảm xúc bình thường có 189 câu, cảm xúc vui có 142 câu, cảm xúc tức có 147 câu và cuối cùng cảm xúc buồn là 122 câu. Các câu tổng hợp bao gồm các câu ngắn biểu lộ cảm xúc như: "*Có lương rồi*", các câu cảm thán "*Thật á*", "*Ôi chúa ơi*", đến các câu dài như "*Không biết thì dựa cột mà nghe hiểu chưa*". Các câu tổng hợp cũng bao gồm các câu có từ cảm thán như "*Ôi dào, người như vậy không thay đổi được đâu*", và các câu không có từ cảm thán như "*Hôm nay chẳng được việc gì cả*".

Các câu tiếng Việt tổng hợp có cảm xúc sẽ được đánh giá thông qua phương pháp đánh giá chủ quan trong đó sử dụng người nghe để đánh giá và phương pháp đánh giá khách quan trong đó sử dụng đo lường phổ tín hiệu để đánh giá.

4. Đánh giá chất lượng tiếng Việt có cảm xúc bằng phương pháp chủ quan

Kết quả đánh giá chủ quan được thể hiện thông qua ma trận nhầm lẫn. Có 25 sinh viên tham gia nghe để đánh giá trực tiếp, các câu nói tổng hợp có cảm xúc được nghe theo trình tự ngẫu nhiên và người nghe lựa chọn kết quả là một trong bốn cảm xúc bình thường, vui, buồn, tức. Trong 600 câu, có 14 câu xuất hiện đầy đủ cả 4 cảm xúc cho cả 2 giọng nam và nữ. Kết quả đánh giá một trong 15 câu này được thể hiện trong Bảng IV và Bảng V.

Hai bảng này là ma trận nhầm lẫn cho giọng nữ và giọng nam với câu "*Ông nói gì thế tôi không hiểu*", trong đó tỉ lệ nhận đúng thấp nhất là 20% (5/25- trong 25 người nghe chỉ có 5 người nghe cho rằng câu này thể hiện cảm xúc vui) cho cảm xúc vui với giọng nữ và 24% (6/15) với giọng nam. Tỷ lệ nhận đúng cao nhất là 96% (24/25), của giọng nữ cho cảm xúc bình thường còn của giọng nam là 92% (23/25) cho cảm xúc tức. Kết quả này cho thấy, sự biểu cảm cho cảm xúc tức ở nam dễ phân biệt hơn ở nữ. Tỉ lệ nhận đúng cảm xúc vui là thấp như phân tích ở [20] khi cảm xúc này thường khó phát hiện nếu không đi kèm với cử chỉ hoặc khuôn mặt hay các từ ngữ biểu cảm rõ rệt.

Bảng VI
MA TRẬN NHẦM LẪN THỐNG KÊ CẢ GIỌNG NAM VÀ GIỌNG NỮ CHO 600 CÂU

	Bình thường	**Tức**	**Vui**	**Buồn**
Bình thường	2855	229	99	1542
Tức	592	1873	1128	82
Vui	791	937	1625	197
Buồn	1151	171	76	1652

Ngoài ra, tỷ lệ nhận đúng chưa cao lắm do người nghe có thể có xu hướng bị ảnh hưởng của nội dung câu nói. Có các câu nói như "*Có lương rồi*" người nghe sẽ có xu hướng cho rằng đây là cảm xúc vui nhiều hơn tức và buồn, hay với câu nói "*Ôi dào, người như vậy không thay đổi được đâu*", người nghe có xu hướng chọn cảm xúc buồn hoặc tức. Với câu "*Chuyển lá thư này cho anh ấy nhé*", người nghe có thể chọn là vui hoặc buồn chứ không lựa chọn cảm xúc tức hay bình thường.

Bảng VI là ma trận nhầm lẫn được thống kê cho cả giọng nam và giọng nữ với 600 câu đã tổng hợp. Thống kê với tất cả các câu cho thấy, kết quả đánh giá đúng tăng hẳn lên. Với giọng nam, tỷ lệ nhận đúng cao hơn giọng nữ ở các cảm xúc tức và vui. Nguyên nhân là do cách biểu diễn cảm xúc tức và vui ở bộ ngữ liệu đều thể hiện ở phần cuối câu nên khi tổng hợp dễ bị nhầm lẫn. Thêm vào đó, cách thể hiện cảm xúc ở giọng nam thường dứt khoát, to hơn giọng nữ dẫn đến kết quả sai nhầm ít hơn. Ngoài ra, kết quả nghiên cứu của [28] cho thấy, việc bổ sung các tham số khác trong miền tần số ngoài các tham số $F0$, thời hạn, cường độ đã làm tăng kết quả nhận dạng cảm xúc. Điều này cũng có nghĩa là khi tổng hợp tiếng Việt có cảm xúc, nên bổ sung các tham số đặc trưng khác nữa vào tham số điều khiển nhằm thể hiện tốt hơn cảm xúc tiếng Việt tổng hợp.

5. Đánh giá chất lượng tiếng Việt tổng hợp có cảm xúc bằng phương pháp khách quan

Bên cạnh việc sử dụng người nghe để đánh giá chủ quan chất lượng của tiếng nói có cảm xúc được tổng hợp, có thể sử dụng phương pháp đánh giá khách quan chất lượng

này thông qua đo lường khoảng cách phổ giữa tiếng nói tự nhiên và tiếng nói tổng hợp. Các câu tổng hợp với các cảm xúc vui, buồn, tức, bình thường được so sánh phổ với các câu nói tự nhiên tương ứng thể hiện cùng nội dung và cảm xúc đó. Do việc tổng hợp với cảm xúc bình thường chính là giọng trần thuật đã được đánh giá ở mục II-3, nên trong phần này chỉ thực hiện đánh giá cho ba cảm xúc vui, buồn và tức.

Việc đánh giá phổ công suất của khung tín hiệu tiếng nói có thể được thực hiện thông qua các vec tơ đặc trưng. Các hệ số LPC cho phép đánh giá phổ đã được làm trơn theo cách như sau. Theo mô hình tự hồi quy (Auto Recursive) hay mô hình toàn điểm cực, quan hệ đầu vào – đầu ra của mô hình này là

$$x(n) + \sum_{i=1}^{p} a_i x(n-1) = \sigma u(n), \tag{6}$$

trong đó $x(n)$ là tín hiệu tiếng nói bức xạ tại môi, $u(n)$ là tín hiệu bậc đơn vị, σ là độ tăng ích của mô hình, còn a_i là các hệ số của bộ lọc đảo có hàm truyền

$$A(z) = \sum_{i=1}^{p} a_i z^{-i}. \tag{7}$$

Nếu cho tín hiệu tiếng nói qua bộ lọc đảo, tín hiệu ra của bộ lọc này sẽ là sai số tiên đoán LPC và ta có

$$S(\omega)\,|A(\omega)|^2 = \sigma_\omega^2, \tag{8}$$

với $S(\omega)$ là phổ công suất của tín hiệu tiếng nói.

Từ đó có thể thấy rằng, phổ công suất của tín hiệu tiếng nói có thể được xấp xỉ thông qua đáp ứng của bộ lọc toàn điểm cực mà hàm truyền đạt của nó dược lựa chọn sao cho sai số tiên đoán là nhỏ nhất [2]. Như vậy, phổ công suất của khung tín hiệu tiếng nói có thể được đánh giá thông qua các hệ số LPC theo công thức

$$S(\omega) \approx \frac{\sigma_\omega^2}{\left|1 - \sum_{i=1}^{p} a_i e^{-j\omega i}\right|^2}. \tag{9}$$

Khoảng cách phổ dựa trên cách đánh giá trên chỉ giữ lại đặc trưng phổ đã được làm trơn. Phương pháp này có tên gọi là đánh giá log phổ theo trị hiệu dụng (RMS: Root Mean Square).

Giả sử S_{xx} và S'_{xx} là phổ của hai mô hình khác nhau, sai số giữa các phổ này được định nghĩa [38]:

$$E(\omega) = \ln S_{xx} - \ln S'_{xx}. \tag{10}$$

Để đo lường khoảng cách giữa các phổ này, tập các chuẩn L_p đã được định nghĩa bởi d_p [38]:

$$(d_p)^p = \int_{-\pi}^{\pi} |E(\omega)|^p \, \frac{d\omega}{2\pi}. \tag{11}$$

Đánh giá log phổ theo trị hiệu dụng được xác định với $p = 2$. Chuẩn L_p được đánh giá theo mô hình phổ đã làm trơn được xác định từ các hệ số LPC và đánh giá theo công thức (9).

Hình 9 và Hình 10 là độ lệch phổ của tín hiệu tiếng nói tổng hợp và tiếng nói tự nhiên tương ứng cho 14 câu. Trên đồ thị, trục hoành là mã của 14 câu tổng hợp với các cột biểu thị giá trị thấp nhất, cao nhất và trung bình của độ lệch phổ; trục tung là giá trị phần trăm của độ lệch phổ giữa tín hiệu tiếng nói tự nhiên và tiếng nói tổng hợp.

Kết quả cho thấy, phổ của tín hiệu tiếng nói tổng hợp có độ chênh lệch không đáng kể so với phổ của tín hiệu tiếng nói tự nhiên, chênh lệch lớn nhất là 3,63% với giọng nữ cho câu 3302 với cảm xúc buồn, nhỏ nhất là 0,21% cho câu 3201 với cảm xúc buồn. Đối với giọng nam, phổ lệch nhất là 3,63% với cảm xúc buồn cho câu 4102 và thấp nhất là 0,22% cho cảm xúc vui của câu 0104. Điều này cho thấy, bộ tổng hợp tiếng Việt có cảm xúc cho kết quả rất gần với giọng tự nhiên.

Kết quả so sánh phổ cũng cho thấy tính đúng đắn và khả thi về mặt công nghệ khi tín hiệu tổng hợp gần với tín hiệu tiếng nói tự nhiên tương ứng.

IV. KẾT LUẬN

Trong bài báo này, tiếng Việt có cảm xúc đã được tổng hợp bằng phương pháp ghép nối bước đầu cho kết quả khá tốt. Các cảm xúc được thể hiện tốt là cảm xúc tức, buồn và bình thường; cảm xúc vui hay bị nhầm lẫn với cảm xúc bình thường hay tức. Điều này thể hiện đúng với nghiên cứu trong [20] do đặc điểm của cảm xúc này thường đi kèm với biểu cảm của khuôn mặt và hành động, cử chỉ khác chứ không thể hiện nhiều qua giọng nói. Với các cảm xúc mạnh, việc biểu lộ ở giọng nam tốt hơn ở giọng nữ. Tỷ lệ nghe nhầm cao có thể do tham số tiếng nói tổng hợp chưa đạt ở mức hoàn hảo.

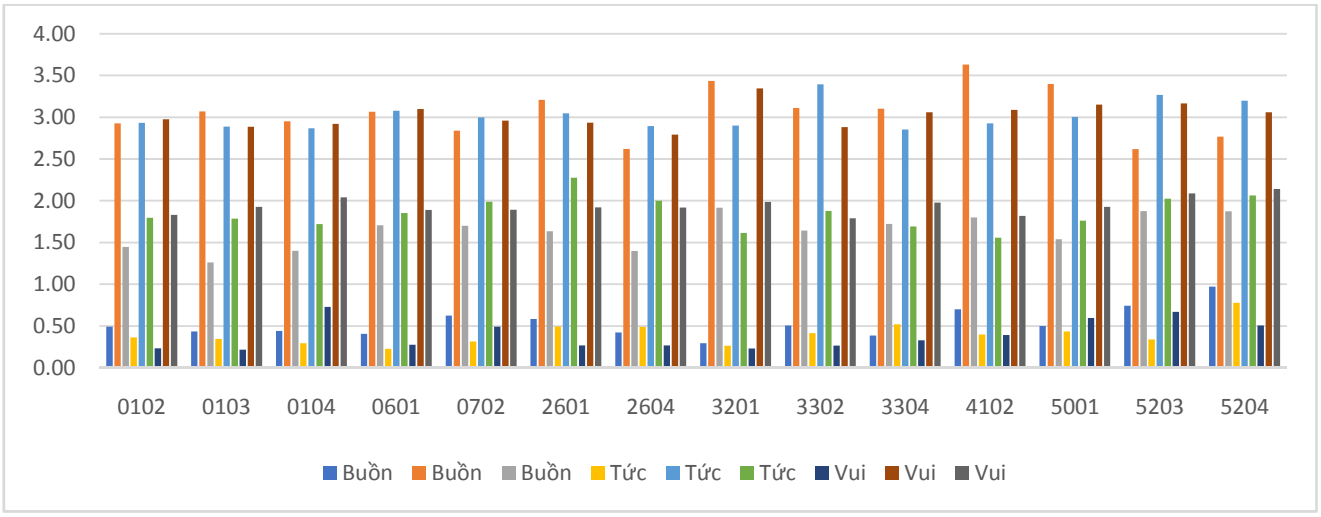
Tiếp theo, nhóm nghiên cứu sẽ cải thiện hơn nữa chất lượng tiếng nói tổng hợp theo phương pháp ghép nối đồng thời xây dựng thêm các bộ ngữ liệu cảm xúc tiếng Việt theo lứa tuổi, vùng miền và mở rộng thêm nghiên cứu với các cảm xúc cơ bản khác.

LỜI CẢM ƠN

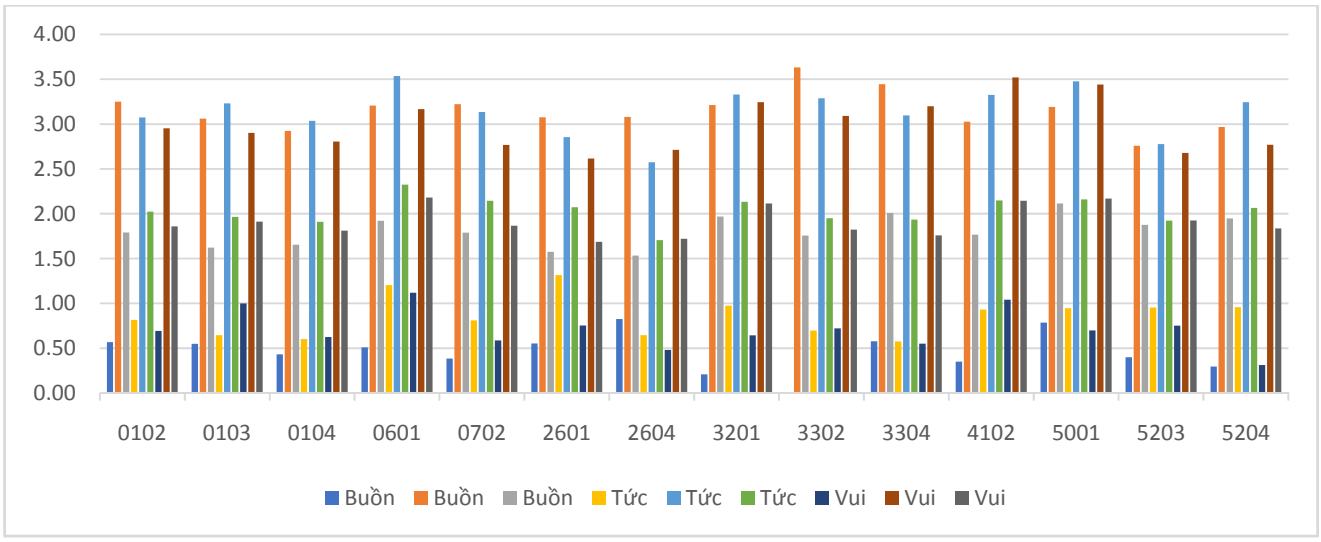
Bài báo này được thực hiện trong khuôn khổ đề tài nghiên cứu "Xây dựng bộ tổng hợp tiếng Việt có cảm xúc", mã số đề tài T2017-PC-077 của Trường Đại học Bách khoa Hà Nội. Các tác giả chân thành cảm ơn Trường Đại học Bách khoa Hà Nội, Phòng Khoa học Công nghệ, Viện Công nghệ Thông tin và Truyền thông, Bộ môn và Phòng Thí nghiệm Kỹ thuật Máy tính đã hỗ trợ để có thể thực hiện thành công đề tài.

TÀI LIỆU THAM KHẢO

[1] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, 1997, vol. 3.



Hình 9. Độ lệch phổ giữa câu tự nhiên và câu tổng hợp cho giọng nam.



Hình 10. Độ lệch phổ giữa câu tự nhiên và câu tổng hợp cho giọng nữ.

[2] J. Holmes and W. Holmes, *Speech Synthesis and Recognition*. Taylor & Francis, Inc., 2001.

[3] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proceedings of the Fourth International Conference on Spoken Language (ICSLP 96)*, vol. 3. IEEE, 1996, pp. 1970–1973.

[4] H. Kellerman, *Emotion: Theory, Research and Experience - Vol. 1*. Academic Press, 1989.

[5] C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustical correlates," *The Journal of the Acoustical Society of America*, vol. 52, no. 4B, pp. 1238–1250, 1972.

[6] F. Burkhardt and W. F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant-synthesis," in *Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[7] S. J. Mozziconacci and D. J. Hermes, "Role of intonation patterns in conveying emotion in speech," in *Proceedings of the 14th International Congress of Phonetic Sciences*, 1999, pp. 2001–2004.

[8] C. Gobl, E. Bennett, and A. N. Chasaide, "Expressive synthesis: how crucial is voice quality?" in *Proceedings of the Workshop on Speech Synthesis*. IEEE, 2002, pp. 91–94.

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

[10] D. T. Nguyen, C. M. Luong, B. K. Vu, H. Mixdorff, and H. H. Ngo, "Fujisaki model based f0 contours in vietnamese tts." in *Proceedings of the 8th International Conference on Spoken Language Processing*, 2004.

[11] A.-T. Dinh, T.-S. Phan, T.-T. Vu, and C. M. Luong, "Vietnamese HMM-based speech synthesis with prosody information," in *Proceedings of the Eighth ISCA Workshop on Speech Synthesis*, 2013.

[12] D. D. Tran, E. Castelli, X. H. Le, J.-F. Serignat, and V. L. Trinh, "Linear F0 contour model for Vietnamese tones and Vietnamese syllable synthesis with TD-PSOLA," in *Tonal Aspects of Languages*, 2006.
[13] D. D. Tran and E. Castelli, "Generation of F0 contours for Vietnamese speech synthesis," in *Proceedings of the International Conference on Communications and Electronics*, 2010, pp. 158–162.
[14] T. T. T. Nguyen, C. d'Alessandro, A. Rilliard, and D. D. Tran, "HMM-based TTS for Hanoi Vietnamese: issues in design and evaluation," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, vol. 13, 2013, pp. 2311–2315.
[15] D.-K. Mac and D.-D. Tran, "Modeling Vietnamese Speech Prosody: A Step-by-Step Approach Towards an Expressive Speech Synthesis System," in *Trends and Applications in Knowledge Discovery and Data Mining (PAKDD 2015 Workshops: BigPMA, VLSP, QIMIE, DAEBH)*. Springer, 2015, pp. 273–287.
[16] D.-K. Mac, E. Castelli, and V. Aubergé, "Modeling the Prosody of Vietnamese Attitudes for Expressive Speech Synthesis," in *Proceedings of the Spoken Language Technologies for Under-Resourced Languages*, 2012.
[17] T. D. Ngo and T. D. Bui, "A study on prosody of vietnamese emotional speech," in *Proceedings of the Fourth International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2012, pp. 151–155.
[18] L. Vutuan, H. Cheng-wei, Z. Cheng, and Z. Li, "Emotional Feature Analysis and Recognition from Vietnamese Speech," *Journal of Signal Processing*, vol. 29, no. 10, pp. 1423–1432, 2013.
[19] J. Zhipeng and H. Chengwei, "High-Order Markov Random Fields and Their Applications in Cross-Language Speech Recognition," *Cybernetics and Information Technologies*, vol. 15, no. 4, pp. 50–57, 2015.
[20] L. X. Thành, Đ. T. L. Thủy, T. V. Loan, and N. H. Quang, "Cảm xúc trong tiếng nói và phân tích thống kê ngữ liệu cảm xúc tiếng Việt," *Chuyên san Các công trình Nghiên cứu, Phát triển và Ứng dụng Công nghệ Thông tin; Tạp chí Bưu chính Viễn thông*, pp. 86–98, 2016.
[21] P. Boersma and D. Weenink, "Praat: doing phonetics by computer, Phonetic Sciences," University of Amsterdam, 2009. [Online]. Available: http://www.fon.hum.uva.nl/praat/ [Accessed 15/10/2017]
[22] S. King, L. Wihlborg, and W. Guo, "The Blizzard Challenge 2017," in *Proceedings of the Blizzard Challenge 2017 Workshop*, Stockholm, 2017.
[23] Calliope, *La parole et son traitement automatique*. Masson Paris, 1989.
[24] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'86)*, vol. 11. IEEE, 1986, pp. 2015–2018.
[25] D. O'Shaughnessy, "Linear predictive coding," *IEEE Potentials*, vol. 7, no. 1, pp. 29–32, 1988.
[26] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.
[27] L. X. Thành, Đ. T. L. Thủy, T. V. Loan, and N. H. Quang, "So sánh hiệu năng một số phương pháp nhận dạng cảm xúc tiếng Việt nói," in *Kỷ yếu hội nghị khoa học công nghệ quốc gia lần thứ IX, Nghiên cứu cơ bản và ứng dụng công nghệ thông tin*, 2016.
[28] Đ. T. L. Thủy, T. V. Loan, N. H. Quang, and L. X. Thành, "Ảnh hưởng của đặc trưng phổ tín hiệu tiếng nói đến nhận dạng cảm xúc tiếng Việt," in *Kỷ yếu hội nghị khoa học công nghệ quốc gia lần thứ X, Nghiên cứu cơ bản và ứng dụng công nghệ thông tin*, 2017.
[29] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of the Acoustical Society of Japan (E)*, vol. 5, no. 4, pp. 233–242, 1984.
[30] B. Uslu and H. G. Ilk, "Fujisaki intonation model in Turkish text-to-speech synthesis," in *Proceedings of the Signal Processing and Communications Applications Conference (SIU 2009)*. IEEE, 2009, pp. 844–847.
[31] E. Navas and I. Hernáez, "Modelado de la entonación en euskera utilizando el modelo de fujisaki y árboles de regresión binarios," *Resumenes de las I Jornadas de Tecnologías del Habla*, 2000.
[32] H. Fujisaki, S. Narusawa, S. Ohno, and D. Freitas, "Analysis and modeling of f_0 contours of portuguese utterances based on the command-response model," in *Proceedings of the Eighth European Conference on Speech Communication and Technology*, 2003.
[33] H. Fujisaki, C. Wang, S. Ohno, and W. Gu, "Analysis and synthesis of fundamental frequency contours of standard Chinese using the command–response model," *Speech communication*, vol. 47, no. 1, pp. 59–70, 2005.
[34] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3. IEEE, 2000, pp. 1281–1284.
[35] H. Mixdorff, N. H. Bach, H. Fujisaki, and M. C. Luong, "Quantitative analysis and synthesis of syllabic tones in Vietnamese," in *Proceedings of the Eighth European Conference on Speech Communication and Technology*, 2003.
[36] T. B. Patel and H. A. Patil, "Analysis of natural and synthetic speech using Fujisaki model," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5250–5254.
[37] B. H. Nguyên and N. T. Dũng, "Mô hình Fujisaki và áp dụng trong phân tích thanh điệu tiếng Việt," in *Kỷ yếu Hội thảo Quốc gia lần thứ 6*, 2003.
[38] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.

Lê Xuân Thành sinh năm 1982. Ông nhận bằng Kỹ sư Điện tử - Viễn thông và Thạc sĩ ngành Xử lý Thông tin và Truyền thông vào các năm 2006 và 2009 tại Trường Đại học Bách khoa Hà Nội. Hiện nay, ông là giảng viên và nghiên cứu sinh tại Bộ môn Kỹ thuật Máy tính, Trường Đại học Bách khoa Hà Nội. Lĩnh vực nghiên cứu của ông là xử lý tín hiệu, xử lý tiếng nói và hệ nhúng.

Trịnh Văn Loan tốt nghiệp Trường Đại học Bách khoa Hà Nội năm 1978. Ông nhận bằng Thạc sĩ năm 1988, và nhận bằng Tiến sĩ năm 1992, tại Viện Đại học Bách khoa Quốc gia Grenoble (INPG) Pháp. Hiện nay, ông đang công tác tại Viện Công nghệ Thông tin và Truyền thông, Trường Đại học Bách khoa Hà Nội. Lĩnh vực nghiên cứu của ông là xử lý tín hiệu, xử lý tiếng nói và hệ nhúng.

Đào Thị Lệ Thủy sinh năm 1976. Tác giả tốt nghiệp Học viện Kỹ thuật Quân sự năm 2008. Hiện nay, tác giả đang là giảng viên Khoa Công nghệ Thông tin, Trường Cao đẳng nghề Công nghệ cao Hà Nội và là nghiên cứu sinh tại Viện Công nghệ Thông tin và Truyền thông, Trường Đại học Bách khoa Hà Nội. Lĩnh vực nghiên cứu của tác giả là xử lý tín hiệu, xử lý tiếng nói và công nghệ phần mềm.

Nguyễn Hồng Quang nhận bằng Kỹ sư Công nghệ Thông tin và Thạc sĩ ngành Xử lý Thông tin và Truyền thông vào các năm 2000 và 2004 tại Trường Đại học Bách khoa Hà Nội. Năm 2008, ông nhận bằng Tiến sĩ ngành Công nghệ thông tin tại Trường Đại học Avignon, Pháp. Từ năm 2000 tới nay, ông là giảng viên Viện Công nghệ Thông tin và Truyền thông, Trường Đại học Bách khoa Hà Nội. Hướng nghiên cứu quan tâm của ông là xử lý tiếng nói, học máy và học sâu.

Đinh Đồng Lưỡng nhận bằng Kỹ sư Công nghệ Thông tin và Thạc sĩ ngành Xử lý Thông tin và Truyền thông vào các năm 2001 và 2009 tại Trường Đại học Bách khoa Hà Nội. Năm 2015, ông nhận bằng Tiến sĩ ngành Kỹ thuật Máy tính tại Trường Đại học Kyung Hee, Hàn Quốc. Từ năm 2001 tới nay, ông là giảng viên Khoa Công nghệ Thông tin, Trường Đại học Nha Trang. Hướng nghiên cứu quan tâm của ông là xử lý tiếng nói, máy học, thị giác máy tính và nhận dạng các hoạt động con người.