ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Ngô Thị Duyên

NGHIÊN CỨU MÔ HÌNH NHÂN VẬT ẢO BIỂU CẢM TRÊN KHUÔN MẶT BA CHIỀU NÓI TIẾNG VIỆT

Chuyên ngành: Khoa học máy tính
Mã số: 62.48.01.01

TÓM TẮT LUẬN ÁN TIẾN SĨ CÔNG NGHỆ THÔNG TIN

Hà Nội –2015

Công trình được hoàn thành tại: Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội.

Người hướng dẫn khoa học: PGS.TS. Bùi Thế Duy
GS.TS. Masato Akagi

Phản biện 1: PGS.TS. Hà Hải Nam

Phản biện 2: PGS.TS. Huỳnh Quyết Thắng

Phản biện 3: PGS.TS. Đỗ Năng Toàn

Luận án tiến sĩ được bảo vệ trước hội đồng cấp Đại học Quốc gia chấm luận án tiến sĩ họp tại P212 – E3 Trường Đại học Công nghệ - ĐHQGHN

Vào hồi 09 giờ ngày 30 tháng 12 năm 2015

Có thể tìm hiểu luận án tại:

-Thư viện Quốc gia Việt Nam

-Trung tâm Thông tin – Thư viện, Đại học Quốc gia Hà Nội

CHƯƠNG 1. GIỚI THIỆU

1.1. Đặt vấn đề

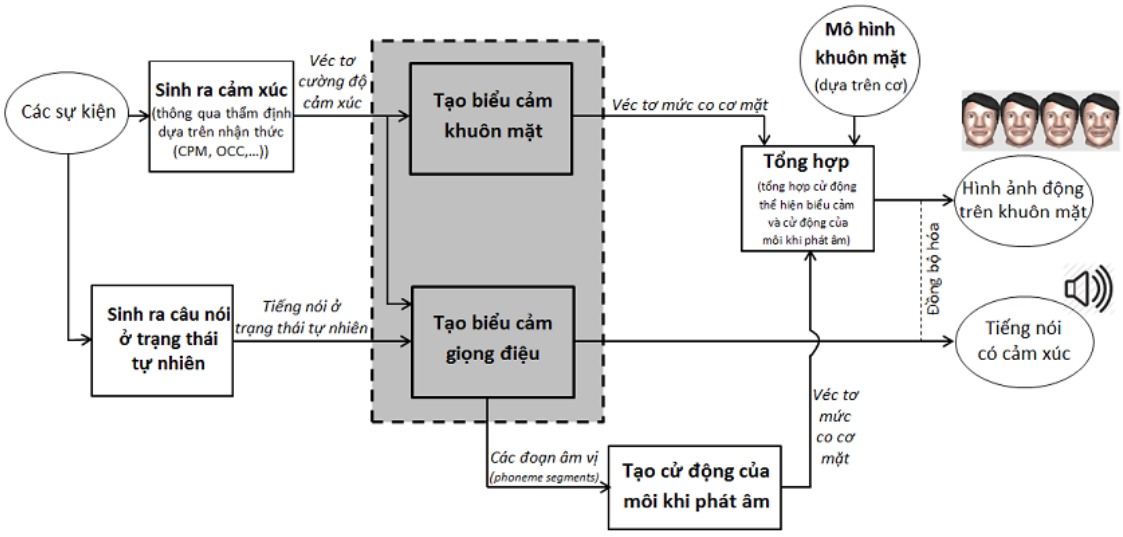
Cùng với sự phát triển nhanh chóng của các lĩnh vực như trí tuệ nhân tạo, đồ họa máy tính, xử lý ngôn ngữ tự nhiên, các nhà nghiên cứu đã giành nhiều công sức hơn nhằm cải tiến sự tương tác giữa người và máy tính, làm cho nó thích hợp, linh động và "hướng con người" hơn. Một phương thức để thực hiện điều đó là thông qua việc tạo các nhân vật ảo. Vì vậy, xây dựng nhân vật ảo là một trong những bài toán đã và đang được quan tâm nhiều bởi miền ứng dụng rộng lớn của chúng: trong giải trí, giáo dục, thương mại điện tử…

Nhân vật ảo là các đối tượng thông minh, có khả năng hoạt động một cách tự chủ, cũng như có các yếu tố giống với con người như cảm xúc, biểu cảm, và hội thoại. Để xây dựng một nhân vật ảo, thông thường chúng ta cần xây dựng ba thành phần sau:

1. Một khuôn mặt có khả năng nói, thể hiện cử động của môi khi nói, thể hiện các biểu cảm và tín hiệu giao tiếp.
2. Một cơ thể có khả năng thể hiện những cử chỉ.
3. Một mô hình trí tuệ bao gồm suy nghĩ, cảm xúc, động lực, hành vi, tính cách... của nhân vật.

Nội dung của luận án nghiên cứu bài toán xây dựng khuôn mặt ba chiều nói tiếng Việt cho nhân vật ảo. Cụ thể, luận án tập trung nghiên cứu một số kỹ thuật thể hiện cảm xúc cho nhân vật ảo nói tiếng Việt.

1.2. Bài toán và cách giải quyết



Hình 1.2: Mô hình cung cấp cảm xúc cho nhân vật ảo.

Nhìn chung, mô hình tổng thể để giải quyết bài toán cung cấp cảm xúc cho nhân vật ảo được thể hiện trên Hình 1.2. Nội dung nghiên cứu của luận án liên quan đến bài toán thể hiện cảm xúc cho nhân vật ảo nói tiếng Việt, liên quan đến các mô đun

nằm trong hình chữ nhật đứt nét trên Hình 1.2. Bài toán thể hiện cảm xúc cho nhân vật ảo có đầu vào là trạng thái cảm xúc liên tục, đầu ra là biểu cảm của nhân vật ảo thể hiện trạng thái cảm xúc đó. Luận án chọn hai kênh biểu cảm là khuôn mặt và tiếng nói để giải quyết bài toán thể hiện cảm xúc cho nhân vật ảo nói tiếng Việt.

Luận án đề xuất ba kết quả nghiên cứu chính góp phần giải quyết bài toán trên.

1. Thứ nhất, để tăng tính tự nhiên, thuyết phục của biểu cảm khuôn mặt thể hiện cảm xúc cho nhân vật ảo, luận án đề xuất mô hình chuyển trạng thái cảm xúc liên tục thành biểu cảm khuôn mặt.
2. Thứ hai, để tạo khả năng thể hiện cảm xúc trong kênh tiếng nói cho nhân vật ảo nói tiếng Việt, luận án đề xuất một mô hình biến đổi tiếng nói tiếng Việt ở trạng thái tự nhiên thành tiếng nói có cảm xúc.
3. Thứ ba, luận án xây dựng một khuôn mặt ba chiều có khả năng thể hiện cảm xúc trên khuôn mặt và trong giọng nói tiếng Việt cho nhân vật ảo. Sau đó, luận án đề xuất phương pháp và tiến hành đánh giá khả năng biểu cảm và độ thuyết phục của khuôn mặt 3D cho nhân vật ảo.

1.3. Cấu trúc của luận án

Ngoài chương Giới thiệu và phần Kết luận, luận án được tổ chức như sau.

Chương 2 trình bày tổng quan các nghiên cứu liên quan đến cảm xúc, mối quan hệ giữa trạng thái cảm xúc và các kênh biểu cảm. Trong chương này, luận án cũng tổng kết các nghiên cứu liên quan tới việc cung cấp cảm xúc và khả năng thể hiện cảm xúc cho nhân vật ảo.

Chương 3 đề xuất mô hình tạo biểu cảm khuôn mặt thể hiện trạng thái cảm xúc liên tục của nhân vật ảo. Mô hình đề xuất thứ nhất dựa trên kết quả nghiên cứu tâm lý và sinh lý học sẽ được trình bày trước. Sau đó luận án đề xuất mô thứ hai dựa trên kết quả phân tích cử động khuôn mặt trong một cơ sở dữ liệu video tự nhiên.

Chương 4 đề xuất một mô hình biến đổi tiếng nói tiếng Việt để thể hiện cảm xúc của nhân vật ảo; mô hình này tổng hợp tiếng nói tiếng Việt có cảm xúc từ đầu vào là tiếng nói ở trạng thái tự nhiên.

Trong **Chương 5,** dựa trên kết quả nghiên cứu được trình bày trong Chương 3 và Chương 4 luận án xây dựng một khuôn mặt ba chiều có khả năng thể hiện trạng thái cảm xúc liên tục một cách tự nhiên trên khuôn mặt cũng như trong giọng nói tiếng Việt.

CHƯƠNG 2. CẢM XÚC VÀ THỂ HIỆN CẢM XÚC CHO NHÂN VẬT ẢO

2.1. Nghiên cứu tâm lý học về cảm xúc

Tổng hợp các nghiên cứu cho thấy có bốn quan điểm chính về mặt tâm lý học để định nghĩa, nghiên cứu, và giải thích về cảm xúc.

Quan điểm Darwin cho rằng cảm xúc là phổ quát và có các chức năng thích nghi. Các nghiên cứu chỉ ra rằng những người đến từ các nền văn hóa khác nhau có thể nhận diện biểu cảm khuôn mặt của một số lượng nhỏ các cảm xúc.

Quan điểm James xem cảm xúc như là các phản ứng của cơ thể, cho rằng những trải nghiệm trong thay đổi của cơ thể chủ yếu bắt nguồn từ trải nghiệm cảm xúc. Ba loại thay đổi cơ thể được xem xét là: hành vi biểu cảm, hành vi công cụ, và những thay đổi sinh lý.

Quan điểm kiến tạo xã hội xem xét cảm xúc như "một vai trò xã hội tạm thời bao gồm đánh giá, thẩm định của cá nhân về tình huống". Quan điểm này tin rằng cảm xúc gắn liền với văn hóa và chỉ có thể được phân tích bằng cách nhìn vào các mức xã hội khác nhau.

Quan điểm nhận thức tin rằng cảm xúc là dựa trên quá trình thẩm định nhận thức. Quan điểm này chỉ ra vai trò của nhận thức trong việc trải nghiệm cảm xúc thông qua việc tập trung vào mối quan hệ giữa cảm xúc và cách mà một người thẩm định các sự kiện trong môi trường. Cảm xúc được xem như là các phản ứng đối với ý nghĩa của sự kiện, liên quan đến mục tiêu và động cơ cá nhân.

2.2. Mối quan hệ giữa cảm xúc và các kênh biểu cảm

2.2.1. Cảm xúc và cử động khuôn mặt

Các nhà nghiên cứu khẳng định rằng tồn tại mối quan hệ giữa cử động khuôn mặt và trạng thái cảm xúc của con người. Hầu hết các nghiên cứu tâm lý học về mối quan hệ giữa cảm xúc và cử động khuôn mặt đi theo một trong ba quan điểm chính:

Quan điểm cảm xúc cơ bản cho rằng có một tập nhỏ các cảm xúc có thể phân biệt hoàn toàn với nhau nhờ biểu cảm khuôn mặt. Theo quan điểm này, tồn tại một mẫu biểu cảm nhất quán, bẩm sinh, và phổ quát cho mỗi cảm xúc cơ bản; trạng thái nào mà không có dấu hiệu khuôn mặt của riêng nó thì không phải là một cảm xúc cơ bản; và tất cả các cảm xúc không phải cảm xúc cơ bản thì đều là sự pha trộn hoặc là nhóm con của các cảm xúc cơ bản.

Quan điểm nhận thức về biểu cảm khuôn mặt thể hiện cảm xúc cho rằng kết quả của quá trình thẩm định gắn liền với những thay đổi trong hoạt động của nhiều hệ thống trong cơ thể, bao gồm cả khuôn mặt.

Quan điểm đa chiều cho rằng các trạng thái cảm xúc về cơ bản được phân biệt dựa trên một số lượng nhỏ các chiều, và rằng cử động khuôn mặt được liên kết với những chiều này.

Hệ mã hóa cử động khuôn mặt (Facial Action Coding System - FACS)

FACS được phát triển nhằm mục đích xác định tất cả các cử động khuôn mặt có thể phân biệt được bằng mắt. FACS liên quan tới việc xác định các cơ mặt khác nhau, hoặc là riêng lẻ, hoặc là theo nhóm gây ra những biến đổi trong hành vi khuôn mặt. Những biến đổi trên khuôn mặt, cùng với cơ bên dưới tạo nên sự biến đổi đó được gọi là các *đơn vị cử động - AU*. FACS là một danh sách gồm 64 đơn vị cử động như vậy. Liên quan đến mối quan hệ giữa cảm xúc và cử động khuôn mặt, mỗi AU mã hóa các cử động cơ bản của một hay một nhóm cơ thường được quan sát thấy khi tạo ra biểu cảm khuôn mặt thể hiện cảm xúc.

2.2.2. Cảm xúc và giọng nói

Tiếng nói là kênh quan trọng thứ hai trong việc thể hiện các trạng thái cảm xúc. Lời nói bao gồm ba thành phần đó là văn phạm, nội dung, và giọng điệu phát âm; trong đó, giọng điệu khi phát âm có ảnh hưởng rất lớn tới việc cảm nhận trạng thái cảm xúc trong hội thoại. Đã có những bằng chứng đáng kể chỉ ra rằng trạng thái cảm xúc có ảnh hưởng trực tiếp tới việc tạo ra phát âm lời nói. Tuy nhiên, cần nhấn mạnh rằng đặc trưng giọng điệu cần phải được đánh giá dựa trên tiêu chuẩn được thiết lập bởi mẫu giọng nói bình thường của một người trong một tình huống cho trước. Trong tất cả các trường hợp, biến đổi của đặc trưng giọng điệu xung quanh một chuẩn sẽ có ý nghĩa hơn là giá trị tuyệt đối.

2.3. Cung cấp cảm xúc cho nhân vật ảo

Có hai vấn đề cần quan tâm khi giải quyết bài toán cung cấp cảm xúc cho nhân vật ảo đó là *cung cấp trạng thái cảm xúc* cho nhân vật ảo và *cung cấp cơ chế thể hiện cảm xúc* cho nhân vật ảo. Đã có những nghiên cứu được đề xuất cho bài toán *cung cấp trạng thái cảm xúc* cho nhân vật ảo. Các mô hình này được đề xuất ở nhiều dạng thức: hệ thống dựa trên luật, hệ thống dựa trên luật mờ, hệ thống phân tán,… Trong số rất nhiều mô hình đã được đề xuất, có rất ít mô hình giải quyết được một cách đầy đủ và thỏa đáng các vấn đề liên quan đến bài toán cái cài đặt cảm xúc trên máy tính, đó là: linh động và độc lập với miền ứng dụng, cảm xúc cần phải có cường độ và cơ chế phân rã theo thời gian, cảm xúc cần phải gắn liền với cá tính và trạng thái động cơ. Mô hình cảm xúc ParleE đề xuất bởi Bui và cộng sự đã giải quyết được các vấn đề này. Với ParleE, nhân vật ảo có khả năng phản ứng lại các sự kiện với cảm xúc thích hợp ở các cường độ khác nhau. Với bài toán *cung cấp cơ chế thể hiện cảm xúc* cho nhân vật ảo, hầu hết các nghiên cứu tập trung vào kênh biểu cảm chính nhất đó là khuôn mặt.

CHƯƠNG 3. MÔ HÌNH THỂ HIỆN CẢM XÚC TRÊN KHUÔN MẶT

3.1. Giới thiệu

Biểu cảm khuôn mặt là một trong những nguồn thông tin quan trọng nhất về trạng thái cảm xúc của một người. Vì vậy, cung cấp cho nhân vật ảo khả năng thể hiện cảm xúc trên khuôn mặt là một trong những yếu tố quan trọng nhằm nâng cao khả năng tương tác của chúng.

3.2. Những nghiên cứu liên quan

Để cung cấp cho nhân vật ảo khả năng thể hiện cảm xúc, trước tiên chúng ta cần hiểu được mối quan hệ giữa cảm xúc và cử động trên khuôn mặt. Theo chúng tôi, trong việc mô phỏng mối quan hệ giữa cảm xúc và cử động khuôn mặt thì các kết quả nghiên cứu thuộc *quan điểm cảm xúc cơ bản* là hữu ích nhất. Cho đến nay, đã có khá nhiều nghiên cứu đi theo *quan điểm cảm xúc cơ bản* để mô phỏng mối quan hệ giữa cảm xúc và khuôn mặt được đề xuất. Tuy nhiên, theo hiểu biết của chúng tôi, hầu như chưa có nghiên cứu nào xem xét động thái theo thời gian của cử động khuôn mặt thể hiện cảm xúc. Ở đây, động thái theo thời gian chỉ thời điểm và khoảng thời gian của các cử động khuôn mặt.

Từ hiểu biết về mối quan hệ giữa cảm xúc và cử động khuôn mặt, nhiều nghiên cứu về thể hiện cảm xúc trên khuôn mặt cho nhân vật ảo đã được đề xuất. Những phương pháp này có thể được chia thành hai lớp:

Phương pháp thể hiện cảm xúc tĩnh: Nhiều nhà nghiên cứu đã sử dụng mô hình bánh xe cảm xúc được mô tả bởi Plutchik để tạo ra các cơ chế ánh xạ trạng thái cảm xúc thành các biểu cảm khuôn mặt được nhận diện một cách phổ biến. Tuy nhiên, mô hình này chỉ là thể hiện cảm xúc tĩnh. Nó không cung cấp một cơ chế nhất quán nào cho việc tạo các biểu cảm cảm xúc trên khuôn mặt. Vì vậy, biểu cảm khuôn mặt bất kỳ có thể được thể hiện ở thời điểm bất kỳ, hoàn toàn độc lập với biểu cảm cảm xúc trước đó của khuôn mặt. Một nhược điểm khác của thể hiện cảm xúc tĩnh là các cảm xúc thường biến đổi tương đối chậm, vì vậy một thay đổi của biểu cảm từ một cảm xúc thành một cảm xúc trái ngược chiếm một thời gian đáng kể, điều này không phù hợp lắm.

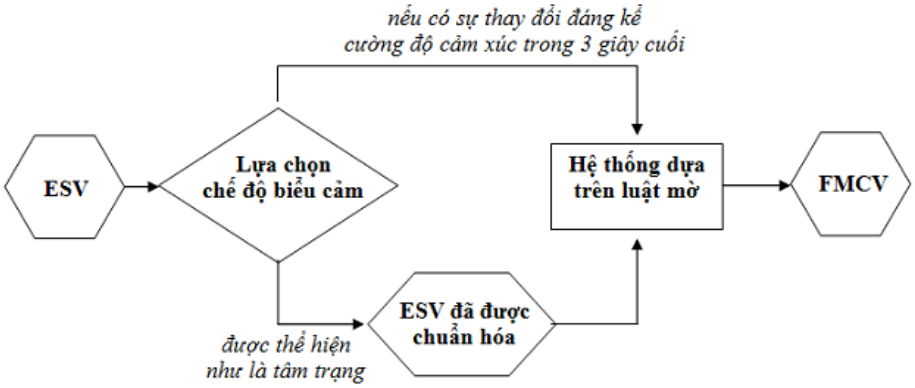
Phương pháp thể hiện cảm xúc động lưu lại sự thay đổi của cường độ cảm xúc theo thời gian, cung cấp một cơ chế nhất quán cho việc tạo biểu cảm khuôn mặt và giải quyết được các giới hạn của phương pháp thể hiện cảm xúc tĩnh. Tuy nhiên, các hệ thống thể hiện cảm xúc động hiện có mới chỉ giải quyết việc thể hiện cảm xúc mà chưa tính đến yếu tố thời gian của các biểu cảm. Trong mỗi khoảng nhỏ thời gian, trạng thái cảm xúc được ánh xạ trực tiếp thành biểu cảm khuôn mặt, sau đó biểu cảm này được thể hiện trên khuôn mặt. Trong trường hợp có một cảm xúc diễn ra trong

một khoảng thời gian dài thì việc ánh xạ trực tiếp từ cảm xúc thành biểu cảm khuôn mặt sẽ làm giảm tính tự nhiên của nhân vật ảo.

3.3. Mô hình tạo biểu cảm khuôn mặt thể hiện trạng thái cảm xúc liên tục

3.3.1. Mô hình đề xuất thứ nhất

Dựa trên kết quả nghiên cứu tâm lý và sinh lý học, luận án đề xuất mô hình thứ nhất tạo biểu cảm khuôn mặt thể hiện trạng thái cảm xúc liên tục của nhân vật ảo. Mô hình này dựa trên ý tưởng rằng một biểu cảm thể hiện cảm xúc trên khuôn mặt xảy ra trong vài giây chỉ khi có sự thay đổi đáng kể trong trạng thái cảm xúc, chính xác hơn là sự tăng đáng kể trong cường độ của các cảm xúc. Khi không có sự thay đổi đáng kể trong cường độ của các cảm xúc, các biểu cảm trên khuôn mặt được giữ ở mức thấp để thể hiện tâm trạng thay vì cảm xúc, ngay cả khi cường độ của cảm xúc là cao. Như chỉ ra trên hình 3.2, mô hình gồm bốn thành phần:



Hình 3.2: Mô hình thứ nhất chuyển trạng thái cảm xúc liên tục thành biểu cảm khuôn mặt

[1.] Đầu vào là chuỗi véc tơ trạng thái cảm xúc (ESV) theo thời gian, kết quả từ một thành phần cảm xúc của nhân vật ảo. Mỗi ESV là một véc tơ chứa cường độ của sáu cảm xúc tại thời điểm *t*, được biểu diễn bởi các số thực:

$$ESV^t = (e^t_1, e^t_2, \ldots, e^t_6) \text{ với } 0 \le e^t_i \le 1.$$

[2.] Đầu ra là chuỗi véc tơ độ co cơ mặt (FMCV) theo thời gian. Mỗi véc tơ FMCV tại thời điểm *t* được mô tả như sau:

$$MFCV^t = (m^t_1, m^t_2, \ldots, m^t_{19}) \text{ với } 0 \le m^t_i \le 1.$$

Đây là một véc tơ biểu diễn mức co của 19 cơ bên phía phải của mô hình khuôn mặt 3D trong mô hình khuôn mặt 3D do Bui đề xuất.

[3.] Mô đun *Lựa chọn chế độ biểu cảm* quyết định một biểu cảm trên khuôn mặt có được tạo ra để thể hiện trạng thái cảm xúc hiện thời hay biểu cảm trên khuôn mặt được giữ ở mức độ thấp để thể hiện tâm trạng thay vì cảm xúc. Thành phần này sẽ thực hiện việc kiểm tra xem có sự tăng đáng kể trong cường độ của cảm xúc bất kỳ kéo dài ba giây (khoảng thời gian của một biểu cảm thể hiện cảm xúc), tức là nếu:

$$e_i^x - e_i^{x-1} > \theta,$$

trong đó $t - 3 \le x \le t$, t là thời điểm hiện tại, và θ là ngưỡng để kích hoạt các biểu cảm thể hiện cảm xúc trên khuôn mặt. Nếu có sự thay đổi đáng kể của cường độ cảm

xúc, véc tơ EVS được chuyển trực tiếp thành véc tơ FMCV dùng *Hệ thống dựa trên luật mờ* được đề xuất bởi Bui. Ngược lại, khi không có sự thay đổi đáng kể của cường độ cảm xúc, véc tơ EVS được chuẩn hóa tới cường độ thấp hơn và sau đó được chuyển thành véc tơ FMCV cũng dùng hệ thống dựa trên luật mờ như trên.

[4.] *Hệ thống dựa trên luật mờ* đề xuất bởi Bui và cộng sự được dùng để chuyển véc tơ ESV thành véc FMCV.

3.3.2. Mô hình đề xuất thứ hai

Mô hình đề xuất thứ hai dựa trên kết quả phân tích cơ sở dữ liệu video về biểu cảm khuôn mặt thể hiện cảm xúc.

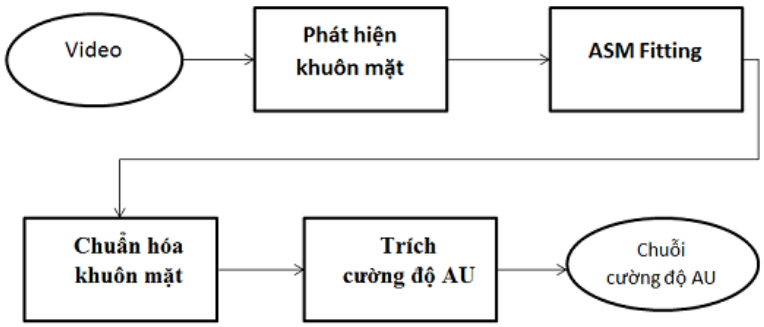
3.3.2.1. Mẫu biểu cảm khuôn mặt thể hiện cảm xúc

Cơ sở dữ liệu

Luận án sử dụng một cơ sở dữ liệu biểu cảm khuôn mặt tự nhiên. Từ ba cơ sở dữ liệu là MMI, FEEDTUM và DISFA, chúng tôi chọn các video trong đó khuôn mặt người tham gia bắt đầu từ trạng thái tự nhiên, tiến dần tới trạng thái đỉnh điểm của biểu cảm, và sau đó trở lại trạng thái tự nhiên. Cuối cùng có 215 video được chọn: vui - 67 video, buồn - 25 video, giận - 25 video, khinh bỉ - 33, sợ hãi - 30 video, và ngạc nhiên - 35 video.

Phân tích cử động khuôn mặt thể hiện cảm xúc

Quá trình phân tích cử động khuôn mặt thể hiện cảm xúc được minh họa trong Hình 3.3.



Hình 3.3: Sơ đồ khối của hệ thống phân tích cử động khuôn mặt thể hiện cảm xúc

A. Mô đun *Phát hiện khuôn mặt*: Với mỗi frame của video đầu vào, mô đun này sẽ trả về kích thước và vị trí xấp xỉ của khuôn mặt được phát hiện. Luận án sử dụng thuật toán Viola Jones để phát hiện khuôn mặt.

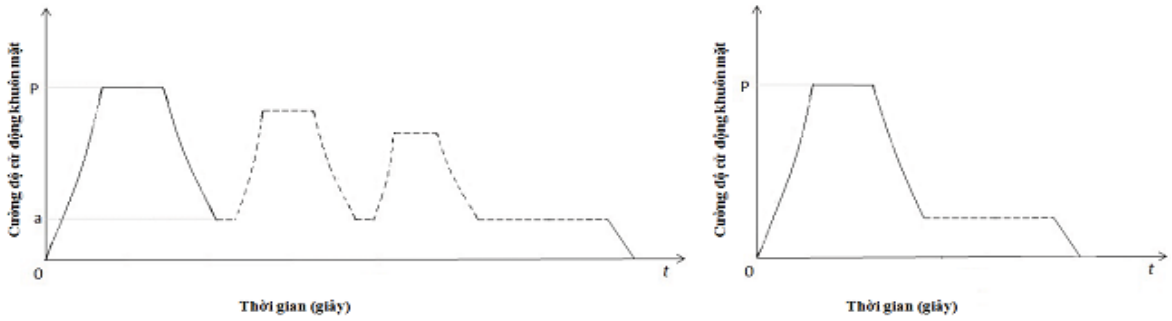
B. Mô đun *ASM Fitting*: Mô đun này dùng thuật toán ASM fitting để trích ra các điểm đặc trưng từ khuôn mặt được phát hiện. Trong vùng khuôn mặt được trả về từ mô đun *Phát hiện khuôn mặt*, mô đun *ASM Fitting* sử dụng Active Shape Model để tìm kiếm vị trí chính xác của các điểm đặc trưng trên khuôn mặt. Đầu ra của mô đun ASM Fitting là vị trí của 68 điểm đặc trưng trên khuôn mặt (ASM shape).

C. Mô đun *Chuẩn hóa khuôn mặt*: Mô đun này dùng khoảng cách giữa hai con ngươi mắt để thực hiện việc chuẩn hóa. Các ASM shape sẽ được chuẩn hóa sao cho khoảng cách giữa hai con ngươi mắt trong các ASM shape là bằng nhau.

D. Mô đun *Trích cường độ AU*: Mô đun này sử dụng các điểm đặc trưng có được từ thao tác chuẩn hóa để trích ra các đặc trưng khuôn mặt liên quan tới sáu cảm xúc cơ bản. Nó dùng vị trí của các điểm đặc trưng đã được chuẩn hóa để tính cường độ của các AU liên quan đến trạng thái cảm xúc được thể hiện trong video đầu vào.

Kết quả phân tích

Từ việc quan sát các đồ thị cường độ AU theo thời gian, chúng tôi đề xuất các mẫu theo thời gian được định nghĩa trước cho biểu cảm khuôn mặt của sáu cảm xúc cơ bản. Mẫu theo thời gian cho biểu cảm thể hiện cảm xúc vui và cảm xúc buồn được mô tả trong Hình 3.6 (a); mẫu theo thời gian cho biểu cảm thể hiện các cảm xúc khinh bỉ, giận, sợ, và ngạc nhiên được chỉ ra trong Hình 3.6 (b).



Hình 3.6 (a): Mẫu theo thời gian của biểu cảm khuôn mặt thể hiện cảm xúc vui và cảm xúc buồn. (b): Mẫu theo thời gian của biểu cảm khuôn mặt thể hiện cảm xúc sợ, giận, ngạc nhiên, khinh bỉ.

Chúng tôi định nghĩa một chu kỳ biểu cảm như sau: $E = (P, Ts, Te, Do, Dr)$ trong đó P là cường độ đích của biểu cảm; Ts và Te là thời gian bắt đầu và thời gian kết thúc của chu kỳ; Do, Dr tương ứng là là khoảng thời gian onset và khoảng thời gian offset của chu kỳ. Quá trình một chu kỳ biểu cảm xuất hiện được mô tả như một hàm theo thời gian:

$$F_e(t) = \begin{cases} P.\emptyset_+(t - Ts, Do) & nếu\ (Ts < t < Ts + Do) \\ P & nếu\ (Ts + Do \le t \le Te - Dr) \\ P.\emptyset_-(t - Te + Dr, Dr) & nếu\ (Te - Dr < t < Te) \end{cases}$$

trong đó $Ø_+$ và $Ø_-$ là các hàm mô tả giai đoạn onset và offset của chu kỳ biểu cảm.

Hàm mô tả phần onset: $\emptyset_+(x, Do) = \exp\left(\frac{ln2}{Do}x\right) - 1$.

Hàm mô tả phần offset:

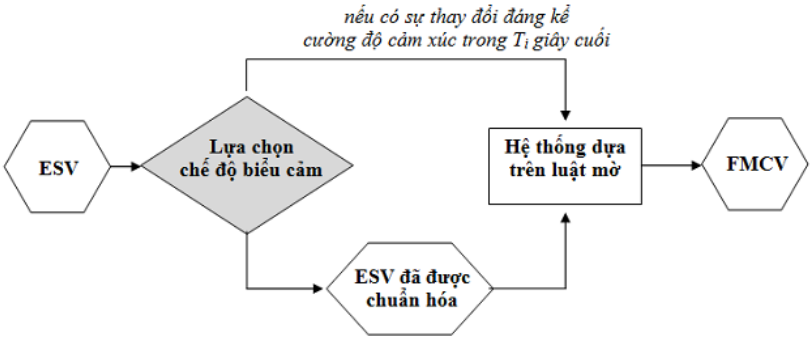
$$\emptyset_-(x, Dr) = \exp\left(ln2 - \frac{ln2 - \ln(\frac{a}{P}+1)}{Dr}x\right) - 1.$$

Để xác thực tính hợp lý của các mẫu theo thời gian được định nghĩa trước, chúng tôi đã thực hiện thao tác so khớp (fitting) cho tất cả các mẫu AU theo thời gian. Thực hiện thao tác so khớp cho tất cả các mẫu AU chúng tôi thu được giá trị trung bình của tổng bình phương lỗi là 0.055 với độ lệch chuẩn là 0.078. Những giá trị này cho thấy mẫu theo thời gian và hàm so khớp ở trên là hợp lý. Kết quả phân tích cho thấy khoảng thời gian trung bình của một chu kỳ biểu cảm cho cảm xúc vui là 3.5 giây,

cho cảm xúc buồn là là 5.3 giây, cho cảm xúc khinh bỉ là 3.6 giây, cho cảm xúc giận và sợ hãi là 3 giây, cho cảm xúc ngạc nghiên là 2.7 giây.

3.3.2.2. Mô hình đề xuất

Các mẫu theo thời gian của cử động khuôn mặt thể hiện các cảm xúc cơ bản được sử dụng làm cơ sở để điều khiển việc tạo biểu cảm khuôn mặt thể hiện cảm xúc. Mô hình đề xuất thứ hai dựa trên ý tưởng rằng khi một cảm xúc được kích hoạt, biểu cảm khuôn mặt sẽ xảy ra theo chuỗi với cường độ giảm dần.



Hình 3.8: Mô hình thứ hai chuyển trạng thái cảm xúc liên tục thành biểu cảm khuôn mặt

Như chỉ ra trên hình 3.8, mô hình đề xuất thứ hai cũng gồm bốn thành phần như mô hình đề xuất thứ nhất. Tuy nhiên, trong mô hình thứ hai, hoạt động của mô đun *Lựa chọn chế độ biểu cảm* được cải tiến so với hoạt động của mô đun này trong mô hình thứ nhất, cụ thể như sau: Nó sẽ thực hiện việc kiểm tra xem có sự tăng đáng kể trong cường độ của cảm xúc bất kỳ trong T_i giây cuối (T_i là khoảng thời gian một chu kỳ biểu cảm), tức là nếu:

$$e_i^x - e_i^{x-1} > \theta,$$

trong đó $t - T_i \leq x \leq t$, t là thời điểm hiện tại, và θ jà ngưỡng để kích hoạt các biểu cảm thể hiện cảm xúc trên khuôn mặt. Nếu có sự thay đổi đáng kể của cường độ cảm xúc, véc tơ EVS được chuyển trực tiếp thành véc tơ FMCV dùng Hệ thống dựa trên luật mờ được đề xuất trong bởi Bui và cộng sự; và thẻ *cycle-tag*$_i$ được đặt giá trị là 1 cho cảm xúc vui và cảm xúc buồn, được đặt giá trị là 3 cho các cảm xúc còn lại. Ngược lại, khi không có sự thay đổi đáng kể của cường độ cảm xúc, véc tơ ESV được chuẩn hóa như sau: Gọi t'_i là thời điểm kết thúc của chu kỳ biểu cảm gần nhất, t là thời điểm hiện tại, khi đó:

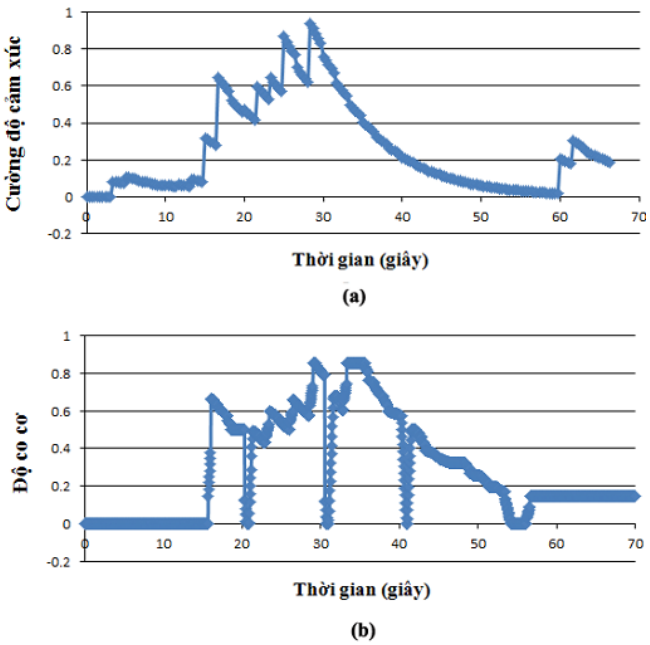
- nếu *cycle-tag*$_i = 1$ và $t'_i + 3 \leq t \leq t'_i + 3 + T_i*0.8$ thì đặt $e^t_i = e^t_i*0.8$ và *cycle-tag*$_i = 2$.
- nếu *cycle-tag*$_i = 2$ và $t'_i + 3 \leq t \leq t'_i + 3 + T_i*0.6$ thì đặt $e^t_i = e^t_i*0.6$ và *cycle-tag*$_i = 3$.
- trường hợp còn lại thì e^t_i được chuẩn hóa về cường độ thấp hơn.

3.4. Thực nghiệm và đánh giá

Luận án sử dụng nhân vật ảo được đề xuất bởi Bui và cộng sự để đánh giá các mô hình tạo biểu cảm khuôn mặt thể hiện trạng thái cảm xúc liên tục được đề xuất. Theo

hiểu biết của chúng tôi, cho tới nay, đây là nhân vật ảo duy nhất có khả năng ánh xạ trạng thái cảm xúc liên tục thành biểu cảm khuôn mặt tương ứng.

Trước khi áp dụng mô hình đề xuất, thỉnh thoảng Obie thể hiện biểu cảm khuôn mặt với cường độ lớn trong một khoảng thời gian dài, khi mà có một cảm xúc nào đó diễn ra trong một thời gian dài. Điều này khiến nhân vật ảo có một diện mạo máy móc, không được tự nhiên. Có thể dễ dàng nhận thấy điều này ở đồ thị trong Hình 3.9.



Hình 3.9: (a): Đồ thị thể hiện cường độ cảm xúc vui của Obie trong trận bóng đá. (b): Đồ thị thể hiện mức co của cơ Zymgomatic Major – cơ cười thể hiện cảm xúc vui trước khi áp dụng mô hình của chúng tôi.

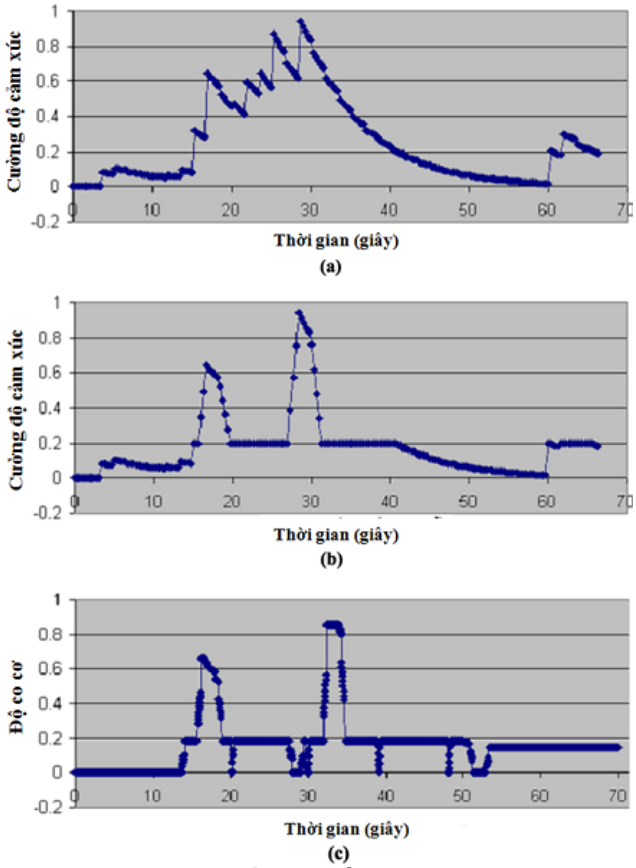
Sau khi áp dụng mô hình đề xuất thứ nhất, mỗi biểu cảm trên khuôn mặt Obie chỉ kéo dài khoảng ba giây. Trong khoảng thời gian còn lại, mặc dù cảm xúc vui vẫn còn tồn tại với cường độ cao, khuôn mặt chỉ thể hiện biểu cảm ở cường độ thấp để diễn tả tâm trạng vui. Có thể nhận thấy điều này từ Hình 3.10 và Hình 3.11.

Sau khi áp dụng mô đề xuất thứ hai, khi cảm xúc vui với cường độ cao xảy ra trong khoảng thời gian dài, biểu cảm trên khuôn mặt Obie chỉ xuất hiện vài chu kỳ với cường độ và khoảng thời gian giảm dần. Trong khoảng thời gian còn lại, mặc dù cảm xúc vui vẫn còn tồn tại với cường độ cao, khuôn mặt chỉ thể hiện biểu cảm ở cường độ thấp để diễn tả tâm trạng vui. Có thể nhận thấy điều này từ Hình 3.12 và Hình 3.13.

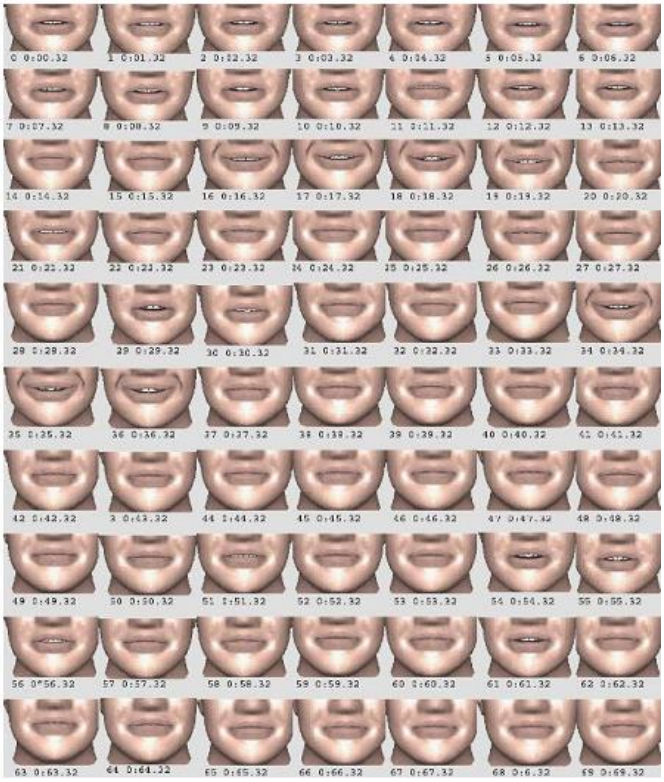
Thực nghiệm đánh giá với người dùng

Quá trình tiến hành thực nghiệm và kết quả đánh giá như sau: Thực nghiệm được tiến hành với ba nhân vật ảo:

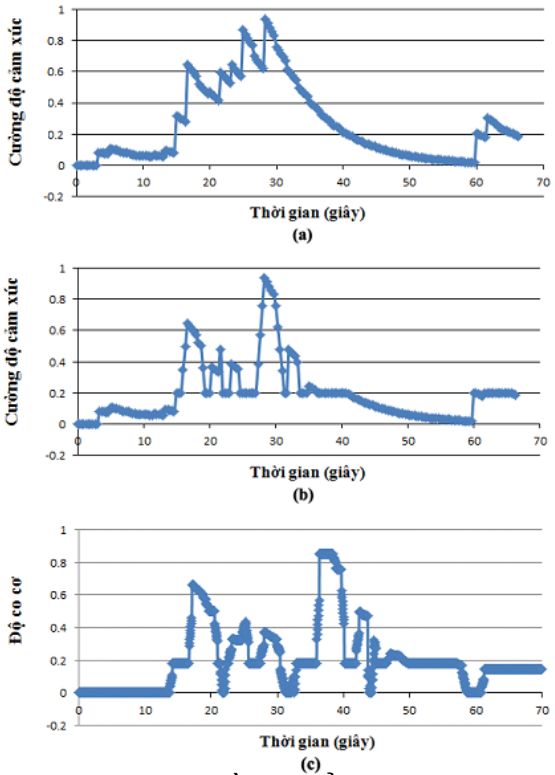
- Nhật vật ảo A: là nhân vật ảo cổ động viên bóng đá Obie nói trên; nhân vật ảo này sử dụng cơ chế ánh xạ trực tiếp để chuyển trạng thái cảm xúc liên tục thành biểu cảm khuôn mặt.
- Nhân vật ảo B: chính là một bản sao của nhân vật ảo A nhưng cơ chế ánh xạ trực tiếp được thay thế bằng mô hình đề xuất thứ nhất.
- Nhân vật ảo C: chính là một bản sao của nhân vật ảo A nhưng cơ chế ánh xạ trực tiếp được thay thế bằng mô hình đề xuất thứ nhất.



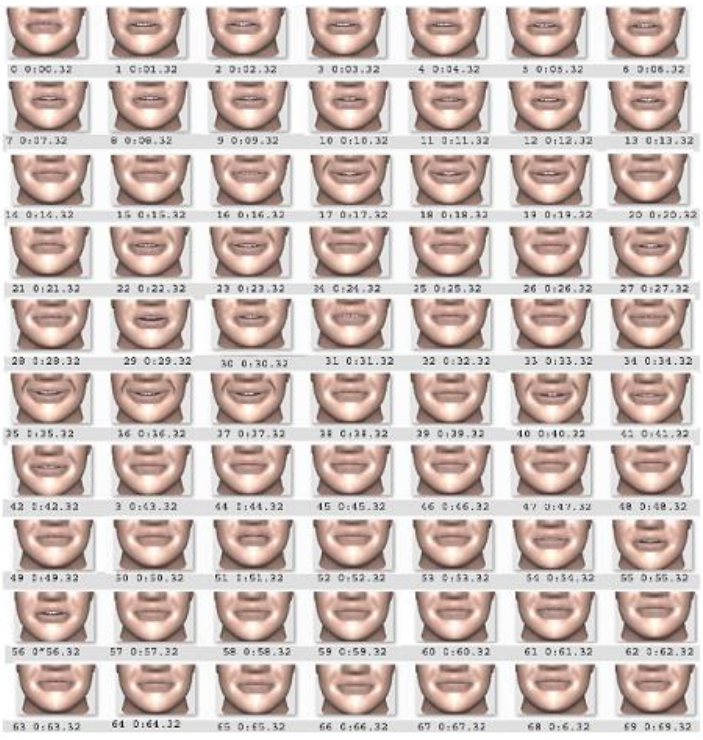
Hình 3.10: (a): Đồ thị thể hiện cường độ cảm xúc vui của Obie trong trận bóng đá. (b): Đồ thị thể hiện cảm xúc vui của Obie được chuẩn hóa bởi mô hình đề xuất thứ nhất. (c): Đồ thị thể hiện mức co của cơ Zymgomatic Major sau khi áp dụng mô hình đề xuất thứ nhất.



Hình 3.11: Biểu cảm khuôn mặt thể hiện cảm xúc vui sau khi áp dụng mô hình đề xuất thứ nhất



Hình 3.12: (a): Đồ thị thể hiện cường độ cảm xúc vui của Obie trong trận bóng đá. (b): Đồ thị thể hiện cảm xúc vui của Obie được chuẩn hóa bởi mô hình đề xuất thứ hai. (c): Đồ thị thể hiện mức co của cơ Zymgomatic Major sau khi áp dụng mô hình đề xuất thứ hai.

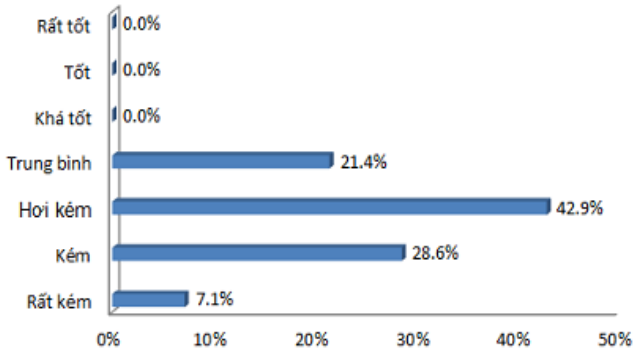


Hình 3.13: Biểu cảm khuôn mặt thể hiện cảm xúc vui sau khi áp dụng mô hình đề xuất thứ hai.

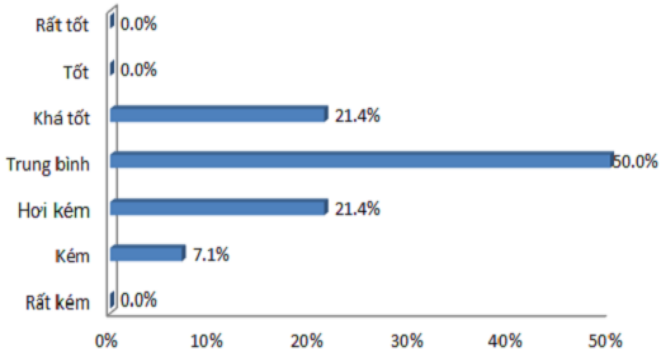
Chúng tôi xây dựng một video clip có hình ảnh gồm hai phần: phần trên là hình ảnh khuôn mặt của ba nhân vật ảo A, B, C được xếp theo thứ tự ngẫu nhiên, phần dưới là hình ảnh thể hiện cường độ theo thời gian của sáu cảm xúc cơ bản mà các nhân vật ảo sẽ thể hiện trên khuôn mặt. Người tham gia thực nghiệm sẽ đánh giá tính thuyết phục trong việc thể hiện cảm xúc trên khuôn mặt của mỗi nhân vật ảo theo thang điểm từ 0 đến 6 (0-Rất kém, 1-Kém, 2- Hơi kém, 3-Trung bình, 4-Khá tốt, 5-Tốt, 6-Rất tốt). Thực nghiệm được tiến hành với 14 người tham gia. Sau khi tiến hành thực nghiệm, kết quả đánh giá của người dùng được tổng kết trong Bảng 3.3, Hình 3.16, Hình 3.17, và Hình 3.18. Từ kết quả đánh giá có thể thấy *nhân vật ảo B thuyết phục hơn nhân vật ảo A trong việc tạo biểu cảm thể hiện cảm xúc trên khuôn mặt* (kết luận 1), và *nhân vật ảo nhân vật ảo C thuyết phục hơn nhân vật ảo B trong việc tạo biểu cảm thể hiện cảm xúc trên khuôn mặt* (kết luận 2). Dùng kết quả trong Bảng 3.3, chúng tôi tiến hành thực hiện kiểm định thống kê để xác thực tính đúng đắn của hai kết luận này.

Bảng 3.3: Tóm tắt kết quả đánh giá tính thuyết phục của các nhân vật ảo trong việc tạo biểu cảm khuôn mặt.

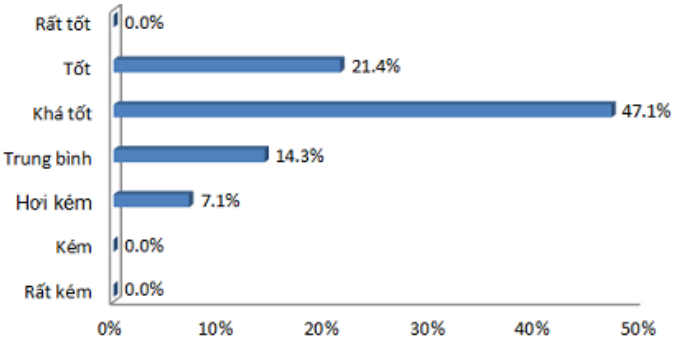
STT	Nhân vật ảo A	Nhân vật ảo B	Nhân vật ảo C
1	2	4	5
2	1	3	4
3	2	4	4
4	2	3	5
5	1	2	3
6	1	4	5
7	3	3	4
8	2	3	4
9	1	1	2
10	2	3	4
11	3	3	4
12	3	2	3
13	2	3	4
14	0	2	4
Trung bình	**1.786**	**2.857**	**3.929**



Hình 3.16: Kết quả đánh giá tính thuyết phục trong việc tạo biểu cảm khuôn mặt của nhân vật ảo A.



Hình 3.17: Kết quả đánh giá tính thuyết phục trong việc tạo biểu cảm khuôn mặt của nhân vật ảo B.



Hình 3.18: Kết quả đánh giá tính thuyết phục trong việc tạo biểu cảm khuôn mặt của nhân vật ảo C.

Kết luận 1: Nhân vật ảo B thuyết phục hơn nhân vật ảo A trong việc tạo biểu cảm thể hiện cảm xúc trên khuôn mặt.

Xét cặp giả thuyết, đối thuyết: H_0: $\mu_A - \mu_B \geq 0$; H_1: $\mu_A - \mu_B < 0$.

Chúng tôi chọn mức ý nghĩa là 0.05 và sử dụng phương pháp kiểm định *matched-pairs t-test*.

Từ kết quả trong Bảng 3.3 sẽ tính được *t = -3.74102*.

Từ giá trị *t* ở trên ta có *P = 0.00123*.

Vì *P = 0.00123 < 0.05* nên giả thuyết H_0 bị từ chối; kết luận *Nhân vật ảo B thuyết phục hơn nhân vật ảo A trong việc tạo biểu cảm thể hiện cảm xúc trên khuôn mặt* được chấp nhận.

Kết luận 2: Nhân vật ảo C thuyết phục hơn nhân vật ảo B trong việc tạo biểu cảm thể hiện cảm xúc trên khuôn mặt.

Xét cặp giả thuyết, đối thuyết: H_0: $\mu_B - \mu_C \geq 0$; H_1: $\mu_B - \mu_C < 0$.

Chúng tôi chọn mức ý nghĩa là 0.05 và sử dụng phương pháp kiểm định *matched-pairs t-test*.

Từ kết quả trong Bảng 3.3 tính được: *t = -8.44639*.

Từ đó có *P = 0.00000*.

Vì *P = 0.00000 < 0.05* nên giả thuyết H_0 bị từ chối; kết luận *Nhân vật ảo C thuyết phục hơn nhân vật ảo B trong việc tạo biểu cảm thể hiện cảm xúc trên khuôn mặt* được chấp nhận.

Từ đây, có thể kết luận nhân vật ảo C (sử dụng mô hình đề xuất thứ hai) thuyết phục nhất (trong A, B, C) trong việc tạo biểu cảm thể hiện cảm xúc trên khuôn mặt.

3.5. Kết chương

Luận án đã đề xuất hai mô hình tạo biểu cảm khuôn mặt thể hiện trạng thái cảm xúc liên tục cho nhân vật ảo. Thực nghiệm đánh giá cho thấy cả hai mô hình đề xuất đều thuyết phục hơn các nghiên cứu trước đó trong việc tạo biểu cảm khuôn mặt thể hiện cảm xúc. Và mô hình đề xuất thứ hai có tính thuyết phục cao hơn, ý tưởng chính là khi một cảm xúc được kích hoạt, biểu cảm khuôn mặt sẽ xảy ra theo chuỗi với cường độ giảm dần và sau đó được giữ ở cường độ thấp để thể hiện tâm trạng, ngay cả khi cảm xúc còn tồn tại ở cường độ cao. Luận án chọn mô hình đề xuất thứ hai khi xây dựng khuôn mặt 3D nói tiếng Việt cho nhân vật ảo.

Kết quả nghiên cứu nêu trên được công bố tại kỷ yếu có phản biện của Hội nghị quốc tế lần thứ 12 và lần thứ 17 về *Multi-Agent Systems - PRIMA 2009, PRIMA 2014* (công trình khoa học số 1, công trình khoa học số 6), kỷ yếu có phản biện của Hội nghị quốc tế lần thứ 6 về *Knowledge and Systems Engineering - KSE 2014* (công trình khoa học số 5), và Tạp chí *Công nghệ thông tin và truyền thông* (công trình khoa học số 2).

CHƯƠNG 4. MÔ HÌNH THỂ HIỆN CẢM XÚC TRONG GIỌNG NÓI TIẾNG VIỆT

4.1. Giới thiệu

Chương này của luận án đề xuất mô hình tạo biểu cảm giọng điệu để thể hiện cảm xúc trong kênh tiếng nói cho nhân vật ảo nói tiếng Việt. Chúng tôi đưa ra cách thức cho việc tổng hợp bốn trạng thái cảm xúc cơ bản của tiếng nói tiếng Việt, thông qua sử dụng các kỹ thuật biến đổi đặc trưng âm, áp dụng cho các phát âm ở trạng thái tự nhiên.

4.2. Những nghiên cứu liên quan

4.2.1. Các phương pháp tổng hợp tiếng nói có cảm xúc

Các phương pháp tổng hợp tiếng nói có cảm xúc có thể được chia thành ba loại chính: tổng hợp tiếng nói có cảm xúc bằng điều khiển tường minh; tổng hợp tiếng nói có cảm xúc bằng phương pháp phát lại, và tổng hợp tiếng nói có cảm xúc bằng điều khiển không tường minh.

4.2.2. Đặc trưng âm liên quan đến tiếng nói có cảm xúc

Tổng hợp các nghiên cứu trước đây đã chỉ ra rằng có hai loại đặc trưng âm có ảnh hưởng lớn đến trạng thái cảm xúc trong tiếng nói đó là ngôn điệu và âm sắc. Về mặt âm học, các đặc trưng âm được xem là quan trọng đối với ngôn điệu phần lớn được trích ra từ tần số cơ bản (F0), năng lượng, và khoảng thời gian. Âm sắc liên quan tới cảm giác thính giác mà người nghe có được trong khi nghe tiếng nói; nó được thể hiện bởi phổ của tín hiệu tiếng nói. Các tham số được phân tích từ phổ được xem là có liên quan tới âm sắc bao gồm tần số cộng hưởng và hình dạng phổ.

Là ngôn ngữ đơn âm tiết và có thanh điệu, tiếng Việt có những đặc trưng riêng biệt so các ngôn ngữ phương Tây. Cho tới nay, đã có một số nghiên cứu về ngôn điệu và âm sắc của tiếng nói tiếng Việt được đề xuất. Một số nghiên cứu về tổng hợp tiếng nói tiếng Việt cũng được công bố. Tuy nhiên, hầu hết các nghiên cứu tập trung vào tiếng nói tự nhiên; có rất ít nghiên cứu về tiếng nói tiếng Việt có cảm xúc.

4.3. Trích đặc trưng âm liên quan tới tiếng nói tiếng Việt có cảm xúc

4.3.1. Cơ sở dữ liệu

Cơ sở dữ liệu tiếng nói có cảm xúc bao gồm các phát âm tiếng Việt được tạo ra bởi một nghệ sĩ nam và một nghệ sĩ nữ. Họ phải phát âm 19 câu ở năm trạng thái cơ bản: tự nhiên, vui, buồn, hơi giận, rất giận. Vì vậy, cơ sở dữ liệu sẽ có tổng số 190 phát âm. Thực nghiệm kiểm tra cảm nhận của người về cảm xúc trong các phát âm đã được thực hiện với 12 sinh viên; kết quả của thực nghiệm cho thấy nhìn chung tỉ lệ nhận dạng thu được là cao.

4.3.2. Giai đoạn trích đặc trưng âm

Đặc trưng âm liên quan tới ngôn điệu được khảo sát bao gồm tần số cơ bản, năng lượng, và khoảng thời gian. Với âm sắc, tần số cộng hưởng và độ nghiêng phổ được phân tích. Ở mức phát âm của câu, có 14 tham số âm được tính và phân tích để tìm ra mối quan hệ giữa sự biến đổi ngôn điệu, âm sắc với trạng thái cảm xúc trong tiếng nói tiếng Việt. Tần số cơ bản trung bình và năng lượng trung bình của các âm tiết cũng được khảo sát.

Cụ thể, giai đoạn trích chọn đặc trưng được thực hiện như sau: Với mỗi phát âm, trước tiên thông tin F0, thông tin năng lượng được trích ra dùng STRAIGHT. Sau đó, từ các thông tin này, một số tham số âm liên quan tới F0 và năng lượng được tính. Các tham số liên quan tới F0 gồm tần số cao nhất (HP), tần số trung bình (AP), và khoảng tần số (PR); tần số trung bình của các âm tiết cũng được xác định; các tham số liên quan tới năng lượng gồm: năng lượng lớn nhất (HPW), năng lượng trung bình (APW), khoảng năng lượng (PWR), năng lượng trung bình của các âm tiết. Tiếp đến, với khoảng thời gian, đối với mỗi phát âm, thông tin về phân đoạn thời gian trước tiên được xác định bằng tay. Việc xác định bao gồm số âm vị, thời gian (ms), và nguyên âm. Khoảng thời gian của tất cả các âm, cũng như khoảng thời gian của khoảng dừng được xác định bằng tay với sự hỗ trợ một phần của Wavesurfer. Từ đó, các tham số liên quan tới khoảng thời gian được xác định bao gồm: trung bình của khoảng dừng (MPAU), tổng thời gian của phát âm (TL), khoảng thời gian của phụ âm (CL), và tỉ lệ giữa khoảng thời gian của phụ âm và khoảng thời gian của nguyên âm (RCV). Cuối cùng, với phổ tín hiệu tiếng nói, các tần số cộng hưởng (F1, F2, F3) và độ nghiêng phổ (ST) được tính. Phổ thu được bằng cách sử dụng STRAIGHT và ba tần số cộng hưởng F1, F2, F3 được tính với LPC-order 12. Độ nghiêng phổ được tính từ H1-A3 trong đó H1 là mức dB của tần số cộng hưởng đầu tiên còn A3 là mức của họa ba có tần số gần nhất với tần số cộng hưởng thứ 3.

Bảng 4.2: Biến đổi trung bình của các tham số âm của bốn trạng thái cảm xúc so với trạng thái tự nhiên.

		vui	buồn	hơi giận	rất giận			vui	buồn	hơi giận	rất giận
Nam	HP	9.28%	-2.25%	8.60%	15.12%	Nữ	HP	12.23%	-0.66%	9.09%	14.37%
	AP	8.09%	-4.60%	6.17%	15.22%		AP	7.75%	-2.10%	6.99%	13.92%
	PR	31.46%	18.66%	15.05%	32.00%		PR	51.57%	28.53%	-11.51%	48.34%
	APW	11.04%	-3.81%	16.04%	19.74%		APW	17.21%	-4.98%	21.45%	27.72%
	HPW	20.81%	-5.84%	13.90%	10.01%		HPW	7.96%	-6.61%	28.97%	28.86%
	PWR	11.53%	-3.26%	22.19%	23.77%		PWR	12.61%	-8.15%	15.79%	20.36%
	MPAU	-6.46%	66.86%	50.86%	59.80%		MPAU	-3.00%	43.95%	-17.03%	37.86%
	CL	-4.96%	9.47%	-10.36%	-1.15%		CL	-3.15%	22.00%	-2.12%	-0.07%
	RCV	-7.50%	-2.13%	-11.72%	2.84%		RCV	-10.24%	-9.87%	-8.23%	1.57%
	TL	-2.50%	15.23%	0.64%	-12.35%		TL	3.55%	16.92%	2.20%	-5.98%
	F1	2.80%	-3.21%	6.29%	10.26%		F1	9.99%	-13.54%	10.82%	20.23%
	F2	1.38%	1.88%	-4.05%	-1.99%		F2	15.43%	-1.87%	-4.21%	-8.87%
	F3	1.42%	-1.17%	-1.84%	5.29%		F3	2.17%	-2%	-4.23%	1.87%
	ST	-15%	6.50%	7.55%	-57%		ST	-14%	5.33%	6.23%	-43%

Sau khi thực hiện giai đoạn trích chọn đặc trưng trên, với mỗi một trong số 190 phát âm của cơ sở dữ liệu, chúng ta có một tập 14 giá trị tương ứng với 14 tham số âm ở mức phát âm của câu. Từ 190 tập này, với các tham số của mỗi trạng thái cảm xúc, các giá trị hệ số biến đổi so với chuẩn được xác định. Kết quả là chúng ta có 152 tập, mỗi tập chứa 14 giá trị của hệ số biến đổi. Trong đó có 19 tập cho mỗi một trong bốn trạng thái cảm xúc, cho mỗi nghệ sĩ tham gia phát âm. Sau đó, với mỗi gói 19 tập này, nhóm các tập có sự tương đồng trong hệ số biến đổi sẽ được chọn. Cuối cùng, từ cụm được chọn, giá trị trung bình của các hệ số biến đổi tương ứng với 14 tham số của mỗi trạng thái cảm xúc được tính. Các giá trị này được liệt kê trong Bảng 4.2. Bảng 4.3 chỉ ra một số kết quả phân tích định lượng ở mức âm tiết. Trong bảng này, thuật ngữ "Âm tiết đầu" chỉ các âm tiết thuộc từ/cụm từ ở vị trí đầu của câu; thuật ngữ "Âm tiết cuối" chỉ các âm tiết thuộc từ/cụm từ ở vị trí kết thúc câu.

Bảng 4.3: Biến đổi trung bình của các tham số của bốn trạng thái cảm xúc so với trạng thái tự nhiên ở mức âm tiết

			vui	buồn	hơi giận	rất giận
Nam	Âm tiết đầu	F-AP	8.58%	-4.85%	6.23%	15.89%
		F-APW	11.5%	-4.04%	17.34%	21.03%
		F-MD	1.05%	15.53%	0.69%	-15.15%
	Âm tiết cuối	L-AP	10.29%	-6.57%	6.98%	17.22%
		L-APW	12.84%	-6.34%	18.05%	25.76%
		L-MD	14.5%	14.98%	-4.69%	-20.42%
Nữ	Âm tiết đầu	F-AP	8.35%	-2.78%	7.65%	14.56%
		F-APW	17.42%	-5.18%	22.62%	28.98%
		F-MD	2.85%	16.99%	2.27%	-8.37%
	Âm tiết cuối	L-AP	9.05%	-3.04%	8.07%	15.42%
		L-APW	19.23%	-7.38%	24.54%	32.68%
		L-MD	16.84%	16.52%	-3.76%	-22.02%

4.4. Tổng hợp tiếng nói tiếng Việt có cảm xúc

4.4.1. Xây dựng luật biến đổi tiếng nói tiếng Việt tự nhiên thành tiếng nói có cảm xúc

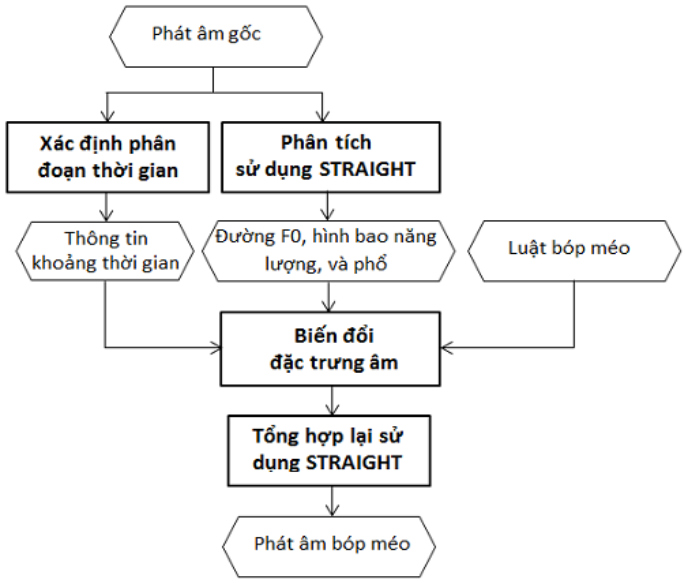
Khi trạng thái cảm xúc trong câu nói tiếng Việt thay đổi, đặc trưng âm không biến đổi đồng đều trong tất cả các âm tiết. Thực tế này đã được xác nhận bởi kết quả phân tích cơ sở dữ liệu trong Phần 4.3.2. Vì vậy, khi xây dựng các luật dùng để biến đổi tiếng nói tiếng Việt tự nhiên thành tiếng nói có cảm xúc, chúng tôi có tính đến sự biến đổi đặc trưng âm ở mức âm tiết. Từ kết quả phân tích được thể hiện trong Bảng 4.2. và Bảng 4.3. có thể xây dựng các luật dùng để biến đổi tiếng nói tiếng Việt tự nhiên thành tiếng nói có cảm xúc, những luật này có tính đến sự biến đổi đặc trưng âm ở mức âm tiết. Ví dụ, luật để tổng hợp cảm xúc vui cho giọng nữ như sau:

(Cảm xúc vui-Giọng nữ){HP:12.23%, AP:7.75%, PR:51.57%, APW:17.21%, HPW:7.96%, PWR:12.61%, MPAU:-3%, CL:-3.15%, RCV:-10.24%, TL:-3.55%, F1:9.99%, F2:15.43%, F3:2.17%, ST:-14%, F-AP:8.35%, F-APW:17.42%, F-MD:2.85%, L-AP:9.05%, L-APW:19.23%, L-MD:16.84%} (1)

Với luật này, đặc trưng âm được biến đổi không đồng đều ở các âm tiết. Ví dụ, khoảng thời gian của các âm tiết thường đều được điều chỉnh giảm, nhưng khoảng thời gian của các âm tiết đầu/cuối phát âm lại được điều chỉnh tăng.

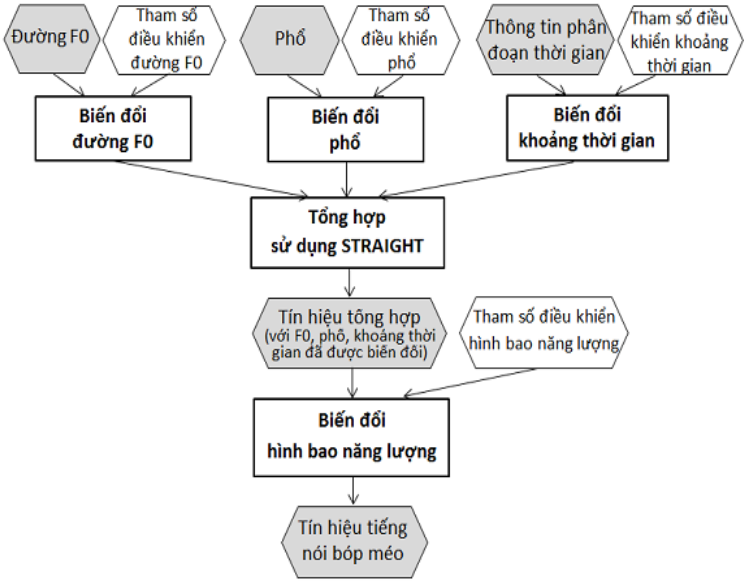
4.4.2. Tiến trình tổng hợp tiếng nói có cảm xúc

Luận án sử dụng kỹ thuật biến đổi tiếng nói để tổng hợp tiếng nói tiếng Việt có cảm xúc từ đầu vào là tiếng nói tự nhiên. Tiến trình thực hiện biến đổi tiếng nói được thể hiện trong Hình 4.2.



Hình 4.2: Tiến trình bóp méo tiếng nói sử dụng STRAIGHT

Trước tiên, STRAIGHT được dùng để trích ra đường F0, hình bao năng lượng, và phổ của tín hiệu tiếng nói tự nhiên, trong khi đó, thông tin phân đoạn thời gian được xác định bằng tay. Sau đó đặc trưng âm liên quan tới F0, năng lượng, phổ, và khoảng thời gian được biến đổi dựa trên các luật biến đổi suy ra từ tập các hệ số trong Bảng 4.2. Quá trình biến đổi này được thực hiện có tính đến sự thay đổi của tham số đặc trưng âm ở mức âm tiết như đã chỉ ra trong Bảng 4.3. Cuối cùng, tiếng nói có cảm xúc được tổng hợp từ đường F0, hình bao năng lượng, phổ, và khoảng thời gian đã được biến đổi thông qua sử dụng STRAIGHT. Quá trình biến đổi được thực hiện theo tiến trình trong Hình 4.3.



Hình 4.3: Tiến trình biến đổi đặc trưng âm

4.5. Thực nghiệm và đánh giá

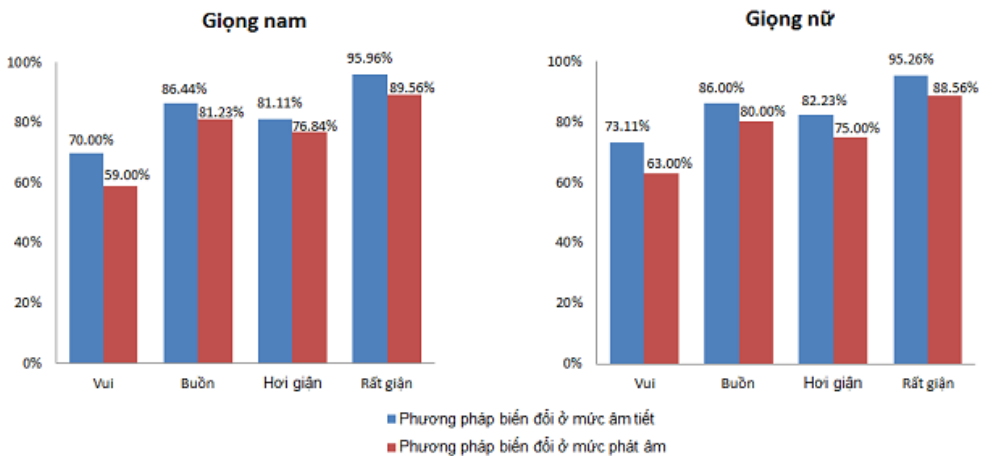
Chúng tôi chọn 10 câu tiếng Việt khác với các câu được sử dụng để trích ra kết quả biến đổi đặc trưng âm trong Phần 4.3. Sau đó, các phát âm ở trạng thái tự nhiên của 10 câu vừa nêu được tạo bởi 1 nam và 1 nữ. Các phát âm ở trạng thái tự nhiên này sẽ được sử dụng để tổng hợp tiếng nói có cảm xúc.

Trước tiên, các luật như được trình bày trong Phần 4.4.1 được áp dụng để tổng hợp tiếng nói có cảm xúc theo tiến trình được trình bày trong Phần 4.4.2. Chúng tôi gọi đây là "Phương pháp biến đổi ở mức âm tiết". Sau đó, chúng tôi cũng tổng hợp tiếng nói có cảm xúc theo tiến trình được trình bày trong Phần 4.4.2 nhưng áp dụng các luật được suy ra **chỉ** từ Bảng 4.2. Chúng tôi gọi đây là "Phương pháp biến đổi ở mức phát âm". Các luật được sử dụng trong "Phương pháp biến đổi ở mức phát âm" không tính đến sự biến đổi đặc trưng âm ở mức âm tiết; với các luật này, đặc trưng âm của các âm tiết được biến đổi đồng đều. Ví dụ, luật tương ứng với luật (1) dùng

để tổng hợp cảm xúc vui cho giọng nữ ở "Phương pháp biến đổi ở mức phát âm" sẽ như sau:

{(Cảm xúc vui-Giọng nữ){HP:12.23%, AP:7.75%, PR:51.57%, APW:17.21%, HPW:7.96%, PWR:12.61%, MPAU:-3%, CL:-3.15%, RCV:-10.24%, TL:-3.55\%, F1:9.99%, F2:15.43%, F3:2.17%, ST:-14%} (2)

Tiếp đến, thực nghiệm đánh giá cảm nhận của người nghe đã được thực hiện cho các phát âm được tổng hợp. Thực nghiệm này được tiến hành theo cách tương tự như thực nghiệm đánh giá trong Phần 4.3.1. Kết quả của thực nghiệm được chỉ ra trên Hình 4.4 cho thấy kết quả nhận dạng tiếng nói tổng hợp của phương pháp biến đổi ở mức âm tiết cao hơn kết quả nhận dạng tiếng nói tổng hợp của phương pháp biến đổi ở mức phát âm; và về mặt tổng thể, kết quả nhận dạng tiếng nói tổng hợp có cảm xúc của phương pháp biến đổi ở mức âm tiết là tương đối cao.



Hình 4.4: Kết quả nhận dạng tiếng nói tổng hợp có cảm xúc

Thực nghiệm đánh giá với người dùng

Thực nghiệm được tiến hành với ba nhân vật ảo:

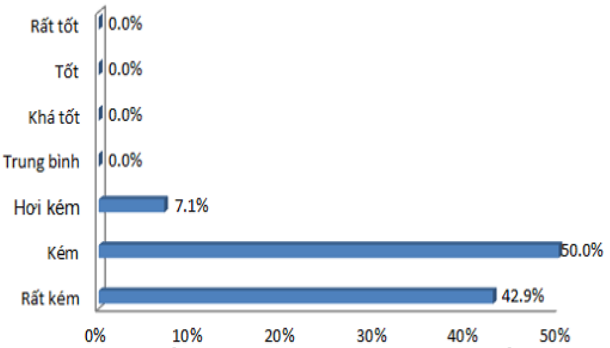
- Nhân vật ảo A: là nhân vật ảo cổ động viên bóng đá Obie nói trên, tiếng nói của nhân vật ảo A là tiếng nói ở trạng thái tự nhiên, không có cảm xúc.
- Nhân vật ảo B: chính là một bản sao của nhân vật ảo A, nhưng ở đây "Phương pháp biến đổi ở mức phát âm" đã được áp dụng để tạo biểu cảm giọng điệu cho nhân vật ảo B.
- Nhân vật ảo C: chính là một bản sao của nhân vật ảo A, nhưng ở đây "Phương pháp biến đổi ở mức âm tiết" đã được áp dụng để tạo biểu cảm giọng điệu cho nhân vật ảo C.

Bảng 4.4: Tóm tắt kết quả đánh giá tính thuyết phục của các nhân vật ảo trong việc tạo biểu cảm giọng điệu.

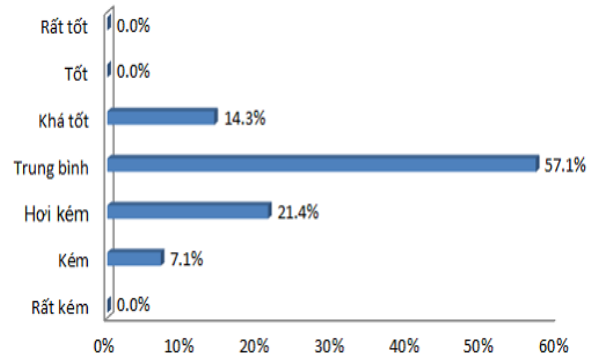
STT	Nhân vật ảo A	Nhân vật ảo B	Nhân vật ảo C
1	0	1	2
2	1	3	4
3	1	4	3
4	1	3	3
5	0	2	3
6	1	2	3
7	2	4	4
8	0	3	3
9	1	3	3
10	1	3	5
11	1	3	4
12	0	3	3
13	0	3	4
14	0	2	4
Trung bình	0.643	2.786	3.429

Mỗi nhân vật ảo được tạo một video clip có hình ảnh gồm hai phần: phần trên là hình ảnh khuôn mặt của nhân vật ảo, phần dưới là hình ảnh thể hiện cường độ theo thời gian của sáu cảm xúc cơ bản mà các nhân vật ảo sẽ thể hiện.

Người tham gia thực nghiệm sẽ đánh giá tính thuyết phục trong việc thể hiện cảm xúc trong giọng nói của mỗi nhân vật ảo theo thang điểm từ 0 đến 6. Thực nghiệm được tiến hành với 14 người tham gia; kết quả đánh giá được tổng kết trong Bảng 4.4, Hình 4.7, Hình 4.8, và Hình 4.9. Từ kết quả đánh giá có thể thấy nhân vật ảo A rất kém trong việc tạo biểu cảm giọng điệu, và bước đầu có thể thấy *nhân vật ảo C thuyết phục hơn nhân vật ảo B trong việc tạo biểu cảm thể hiện cảm xúc trong giọng nói*. Dùng kết quả trong Bảng 4.4 chúng tôi tiến hành thực hiện kiểm định thống kê để xác thực tính đúng đắn của kết luận này.



Hình 4.7: Kết quả đánh giá tính thuyết phục trong việc tạo biểu cảm giọng điệu của nhân vật ảo A.



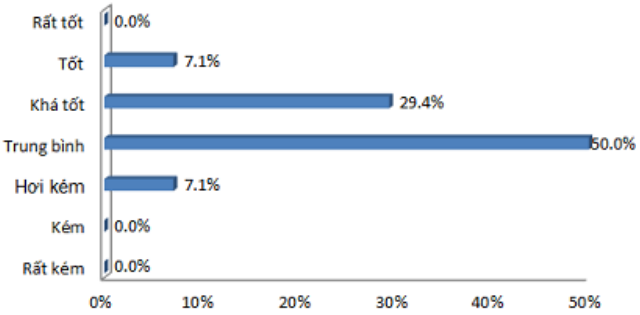
Hình 4.8: Kết quả đánh giá tính thuyết phục trong việc tạo biểu cảm giọng điệu của nhân vật ảo B.

Kết luận: Nhân vật ảo C thuyết phục hơn nhân vật ảo B trong việc tạo biểu cảm thể hiện cảm xúc trong giọng nói.

Xét cặp giả thuyết, đối thuyết:

$$H_0: \mu_B - \mu_C \geq 0 \; ; \; H_1: \mu_B - \mu_C < 0$$

Chúng tôi chọn mức ý nghĩa là 0.05 và sử dụng phương pháp kiểm định *matched-pairs t-test*.



Hình 4.9: Kết quả đánh giá tính thuyết phục trong việc tạo biểu cảm giọng điệu của nhân vật ảo C.

Từ kết quả trong Bảng 4.4 tính được $t = -2.85706$.

Từ giá trị t ở trên ta có $P = 0.00674$.

Vì $P = 0.00674 < 0.05$ nên kết luận trên được chấp nhận. Như vậy, "Phương pháp biến đổi ở mức âm tiết" hiệu quả hơn "Phương pháp biến đổi ở mức phát âm" trong việc tạo biểu cảm giọng điệu cho nhân vật ảo nói tiếng Việt.

4.7. Kết chương

Chương 4 của luận án đã đề xuất mô hình biến đổi tiếng nói tiếng Việt từ trạng thái tự nhiên thành tiếng nói có cảm xúc, cung cấp cho nhân vật ảo khả năng thể hiện cảm xúc trong giọng nói tiếng Việt. Kết quả của thực nghiệm đánh giá cho thấy các trạng thái cảm xúc tổng hợp được nhận dạng tương đối tốt. Kết quả nghiên cứu nêu trên được công bố tại kỷ yếu có phản biện của Hội nghị quốc tế lần thứ 4 và lần thứ 6 về *Knowledge and Systems Engineering - KSE 2012, KSE 2014* (công trình khoa học số 3, công trình khoa học số 4).

CHƯƠNG 5. XÂY DỰNG KHUÔN MẶT BA CHIỀU NÓI TIẾNG VIỆT CHO NHÂN VẬT ẢO

5.1. Giới thiệu

Chương này của luận án mô tả quá trình xây dựng một khuôn mặt ba chiều có khả năng thể hiện cảm xúc trên khuôn mặt và trong tiếng nói cho nhân vật ảo nói tiếng Việt.

5.2. Những nghiên cứu liên quan

Mô hình khuôn mặt ba chiều

Luận án sử dụng mô hình khuôn mặt ba chiều dựa trên cơ được đề xuất bởi tác giả Bui. Mô hình khuôn mặt bao gồm một lưới đa giác thể hiện khuôn mặt, một mặt B-spline thể hiện môi, và một hệ cơ tạo ra biến đổi tự nhiên trên bề mặt khuôn mặt, điều khiển sự tương tác giữa các cơ, và tạo ra các nếp nhăn, điểm lồi, lõm trong thời gian thực.

Tạo chuyển động của môi khi phát âm tiếng nói

Cohen và Masaro đã đề xuất nghiên cứu để mô hình hóa hiệu ứng đồng phát âm trên các chuyển động của môi khi nói. Đồng phát âm là hiệu ứng pha trộn trong đó các âm vị xung quanh sẽ có ảnh hưởng lên âm vị hiện tại. Một chuyển động của môi tương ứng với một phân đoạn tiếng nói được thể hiện như là một phân đoạn hình vị. Mỗi phân đoạn hình vị này có ưu thế khi phát âm; hàm ưu thế xác định mức gần của môi để đạt tới các giá trị đích của hình vị. Sự chồng nhau của các phát âm theo thời gian được tạo ra bởi các hàm ưu thế chồng nhau của các cử động liền kề tương ứng với các lệnh phát âm. Mỗi cử động có một tập các hàm ưu thế, mỗi hàm cho một tham số; trung bình có trọng số của tất cả các hàm ưu thế sẽ tạo ra hình dáng cuối cùng của môi.

Tổng hợp các cử động trên khuôn mặt

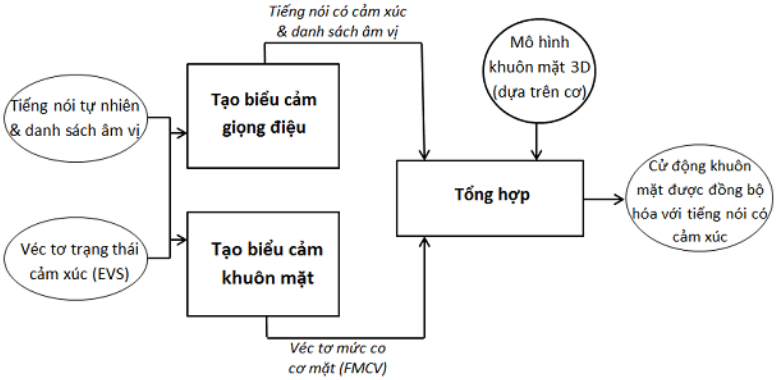
Bui và cộng sự đã đề xuất cơ chế tổng hợp các loại cử động khuôn mặt khác nhau. Cử động khuôn mặt được chia thành các nhóm gọi là các kênh; các tác giả đề xuất một cơ chế để tổng hợp các cử động trong cùng một kênh và một cơ chế để tổng hợp các cử động trong các kênh khác nhau.

Trong cùng một kênh, khi có hai cử động chồng nhau, cử động tổng hợp sẽ đi theo cử động thứ nhất cho tới thời điểm bắt đầu của cử động thứ hai, sau đó cử động tổng hợp sẽ tăng/giảm để tiến tới đích cử cử động thứ hai, và sau đó đi theo cử động thứ hai. Để tổng hợp cử động từ các kênh khác nhau, tác giả đưa ra giải pháp giải quyết vấn đề xung đột giữa các tham số liên quan đến các cử động khác nhau; sau đó, hoạt động của mỗi tham số được tổng hợp bằng cách lấy giá trị lớn nhất của tham số đó từ tất cả các kênh. Tại một thời điểm xác định, khi có xung đột xảy ra giữa các tham số

ở các kênh khác nhau thì tham số liên qua tới cử động với độ ưu tiên cao hơn sẽ chiếm ưu thế và lấn át tham số với độ ưu tiên thấp hơn.

5.3. Kiến trúc hệ thống

Kiến trúc tổng thể của hệ thống khuôn mặt 3D được minh họa trên Hình 5.1. Đầu vào của hệ thống là tiếng nói ở trạng thái tự nhiên cùng với danh sách các âm vị tương ứng có kèm theo thông tin thời gian, và chuỗi các véc tơ trạng thái cảm xúc theo thời gian (EVS).



Hình 5.1: Kiến trúc hệ thống khuôn mặt 3D nói tiếng Việt.

5.3.1. Mô đun *Tạo biểu cảm giọng điệu* (VESS)

Mô đun VESS sử dụng kết quả nghiên cứu đã được trình bày trong Chương 4 để chuyển tiếng nói tiếng Việt ở trạng thái tự nhiên thành tiếng nói có cảm xúc tương ứng với trạng thái cảm xúc đầu vào. Cảm xúc được chọn ở đây là cảm xúc có cường độ cao nhất trong các cảm xúc đầu vào.

5.3.2. Mô đun *Tạo biểu cảm khuôn mặt* (EFE)

Mô đun EFE sử dụng kết quả nghiên cứu đã được trình bày trong Chương 3 để tạo biểu cảm khuôn mặt thể hiện trạng thái cảm xúc liên tục. Đầu vào của mô đun EFE là chuỗi các véc tơ trạng thái cảm xúc (EVS) theo thời gian và đầu ra là chuỗi véc tơ độ co cơ mặt (FMCV) tương ứng.

5.3.3. Mô đun *Tổng hợp*

Mô đun *Tổng hợp* tạo chuyển động của môi khi phát âm tiếng Việt và kết hợp các chuyển động này với cử động khuôn mặt thể hiện cảm xúc.

Hình vị cho các âm vị tiếng Việt: Luận án dựa trên các luật được đưa ra trong nghiên cứu của các tác giả Do và Nguyen để xác định hình vị tương ứng của mỗi âm vị tiếng Việt. Để tạo hình vị cho các nguyên âm, lượng quay của hàm và mức co của các cơ có ảnh hưởng lên môi ban đầu được xác định dựa trên hình thang nguyên âm. Sau đó, những giá trị này được tinh chỉnh lại bằng tay dựa trên sự so sánh giữa hình vị nguyên âm của khuôn mặt 3D với hình vị nguyên âm của khuôn mặt người thật. Để tạo hình vị cho các phụ âm, luận án chia các phụ âm thành ba loại: phụ âm môi - môi, phụ âm môi - răng, và loại thứ ba chứa các phụ âm còn lại. Các luật trong nghiên cứu của tác giả Do và Nguyen được áp dụng để khởi tạo hình vị ban đầu cho các phụ âm; và sau đó các hình vị này cũng được tinh chỉnh lại theo cách tương tự như đã làm cho nguyên âm.

Tổng hợp chuyển động của môi khi phát âm tiếng Việt: Phát âm của một phân đoạn tiếng nói không phải là độc lập, nó phụ thuộc vào các phân đoạn trước và sau

nó. Luận án áp dụng mô hình của tác giả Cohen và Massaro để tạo hiệu ứng đồng phát âm trên các cử động của môi khi phát âm tiếng Việt.

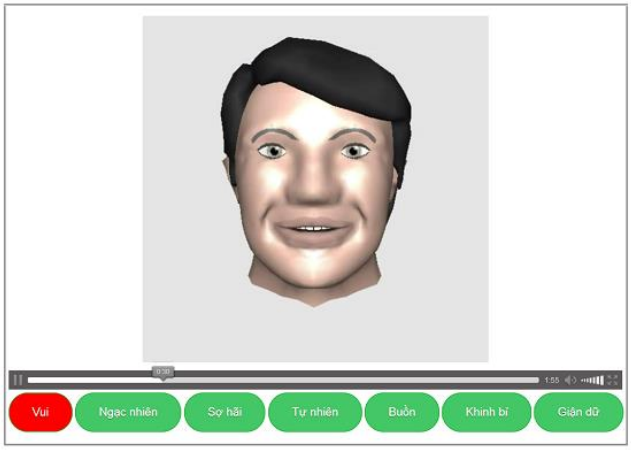
Tổng hợp biểu cảm khuôn mặt và cử động của môi khi phát âm tiếng Việt

Để tổng hợp biểu cảm khuôn mặt thể hiện cảm xúc và cử động của môi khi phát âm tiếng Việt, luận án áp dụng nghiên cứu được đề xuất bởi Bui và cộng sự. Trong hệ thống khuôn mặt nói tiếng Việt, chúng tôi tạo quyền ưu tiên cao hơn cho cử động của môi khi nói. Cử động khuôn mặt cuối cùng, là kết quả của quá trình tổng hợp, sẽ được hiển thị trên khuôn mặt 3D cùng với tiếng nói tổng hợp được đồng bộ hóa.

5.4. Thực nghiệm và đánh giá

Luận án sử dụng ParleE – một mô hình cảm xúc cho nhân vật ảo đề xuất bởi Bui, và đặt khuôn mặt trong miền cổ động viên bóng đá. Mục đích của việc sử dụng ParleE và miền cổ động viên bóng đá là tạo đầu vào để kiểm tra, đánh giá khuôn mặt nói tiếng Việt được xây dựng. Thực nghiệm được tiến hành với hai nhân vật ảo:

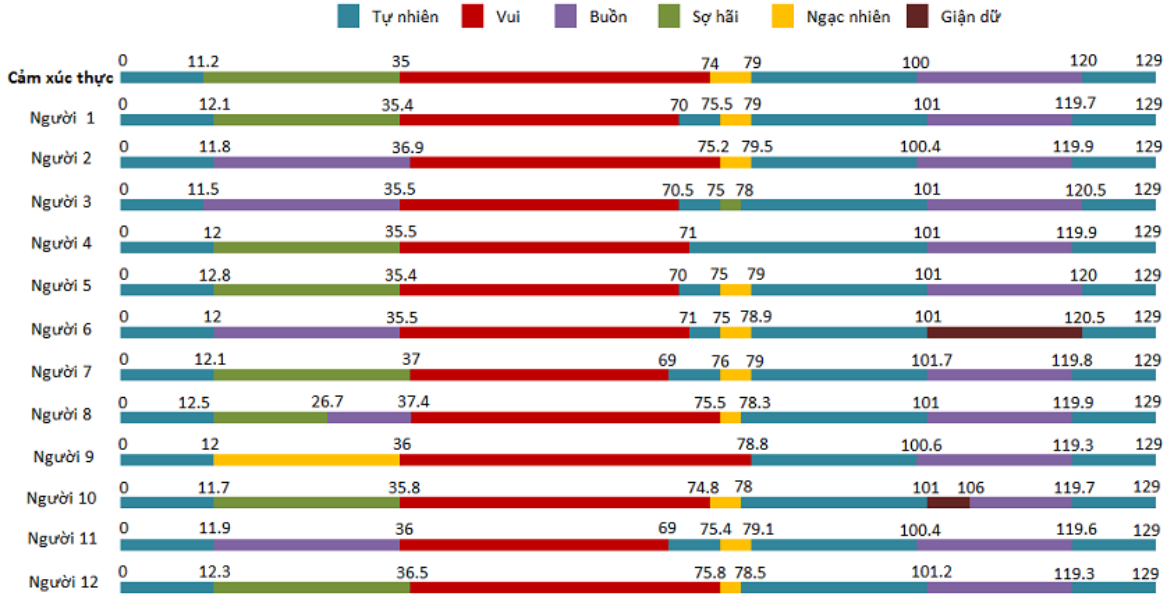
- Nhân vật ảo A: là nhân vật ảo có khuôn mặt ba chiều trong đó mô đun "Tạo biểu cảm giọng điệu" đã bị vô hiệu hóa, nhân vật ảo A chỉ thể hiện cảm xúc trên khuôn mặt, không có tiếng nói.
- Nhân vật ảo B: là nhân vật ảo thể hiện cảm xúc trên cả khuôn mặt và trong giọng nói.



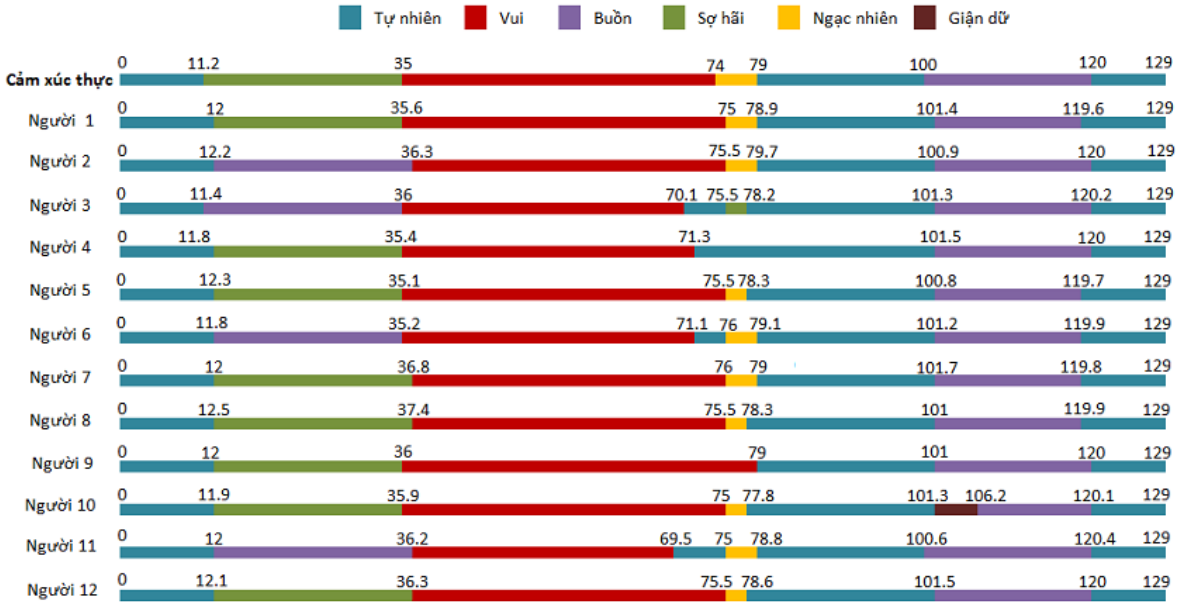
Hình 5.4: Giao diện chương trình ghi lại kết quả cảm nhận của người dùng.

Hai video clip cho hai nhân vật ảo A, B được xây dựng. Mục tiêu của thực nghiệm là ghi lại kết quả cảm nhận trạng thái cảm xúc của người dùng khi xem các video clip, nhằm mục đích so sánh với trạng thái cảm xúc mà thực tế nhân vật ảo cần thể hiện. Một chương trình có giao diện như Hình 5.4 sẽ chạy video clip cho người dùng xem; trong quá trình này, người dùng sẽ chọn trạng thái cảm xúc mà họ nhận thấy nhân vật ảo đang thể hiện bằng cách bấm vào một trong 7 nút bên dưới. Kết quả mà chương trình trả về là các mốc thời gian của từng trạng thái cảm xúc mà người dùng cảm nhận được. Thực nghiệm được tiến hành với 12 người tham gia; kết quả đánh giá được tổng kết trong Hình 5.6 và Hình 5.7. Dòng đầu tiên thể hiện cảm xúc thực mà nhân vật ảo cần phải thể hiện, các dòng tiếp theo thể hiện cảm xúc mà người dùng cảm nhận được từ khuôn mặt của nhân vật ảo. Mỗi cảm xúc được biểu diễn bởi một màu tương ứng; các chỉ số phía trên mỗi dòng là các mốc thời gian tính theo giây. Kết quả đánh giá cho thấy với nhân vật ảo A, khi cảm xúc chỉ được thể hiện trên khuôn mặt mà không có tiếng nói, mặc dù có sự nhầm lẫn hay bỏ sót một số cảm xúc nhưng kết quả cảm nhận của người dùng nhìn chung tương đối tốt. Với nhân vật ảo B, khi cảm xúc được

thể hiện cả trên khuôn mặt và trong giọng nói, kết quả cảm nhận của người dùng khá tốt và tốt hơn so với kết quả cảm nhận của nhân vật ảo A. Như vậy, việc kết hợp thể hiện cảm xúc trên khuôn mặt và trong giọng nói của nhân vật ảo đã làm tăng độ chính xác trong kết quả cảm nhận của người dùng.



Hình 5.6: Kết quả cảm nhận của người dùng về cảm xúc do nhân vật ảo A thể hiện.



Hình 5.7: Kết quả cảm nhận của người dùng về cảm xúc do nhân vật ảo B thể hiện.

5.5. Kết chương

Chương 5 của luận án mô tả quá trình xây dựng một khuôn mặt ba chiều có khả năng thể hiện cảm xúc trên khuôn mặt và trong giọng nói tiếng Việt. Kết quả nghiên cứu nêu trên được công bố tại kỷ yếu có phản biện của Hội nghị quốc tế lần thứ 11 về *Computing and Communication Technologies - RIVF 2015* (công trình khoa học số 7).

KẾT LUẬN

Luận án nghiên cứu bài toán thể hiện cảm xúc cho nhân vật ảo nói tiếng Việt. Luận án đã đề xuất ba kết quả nghiên cứu chính như sau.

Thứ nhất, luận án đề xuất mô hình tạo biểu cảm khuôn mặt thể hiện trạng thái cảm xúc liên tục của nhân vật ảo. Ý tưởng chính của mô hình là khi một cảm xúc được kích hoạt, biểu cảm khuôn mặt sẽ xuất hiện theo chuỗi với cường độ giảm dần. Ý tưởng này xuất phát từ quá trình sử dụng các kỹ thuật nhận dạng biểu cảm khuôn mặt để tự động phân tích một cơ sở dữ liệu video tự nhiên. Kết quả thực nghiêm đánh giá cho thấy mô hình đề xuất góp phần nâng cao tính thuyết phục của nhân vật ảo khi thể hiện cảm xúc trên khuôn mặt.

Thứ hai, luận án đã đề xuất mô hình tạo biểu cảm giọng điệu trong giọng nói tiếng Việt. Từ quá trình phân tích cơ sở dữ liệu tiếng nói tiếng Việt có cảm xúc, các luật thể hiện mối quan hệ về đặc trưng âm giữa tiếng nói có cảm xúc và tiếng nói ở trạng thái tự nhiên được xây dựng. Sau đó, các luật này được sử dụng để biến đổi tiếng nói tiếng Việt ở trạng thái tự nhiên thành tiếng nói tổng hợp có cảm xúc. Kết quả thực nghiệm đánh giá cho thấy tiếng nói tổng hợp được nhận dạng cảm xúc khá tốt.

Thứ ba, luận án đã xây dựng một khuôn mặt ba chiều có khả năng thể hiện cảm xúc trong giọng nói tiếng Việt, đồng thời có khả năng thể hiện cảm xúc trên khuôn mặt cũng như thể hiện cử động của môi khi phát âm các từ tiếng Việt. Khuôn mặt ba chiều này có thể được sử dụng cho các nhân vật ảo nói tiếng Việt, góp phần làm tăng tính tự nhiên, thuyết phục của chúng.

Mặc dù các mô hình đề xuất đã góp phần làm tăng tính thuyết phục của nhân vật ảo trong việc thể hiện cảm xúc. Tuy nhiên, các mô hình này vẫn còn hạn chế là chưa xem xét sự ảnh hưởng của các yếu tố như cá tính, động cơ,... của nhân vật ảo đối với việc thể hiện cảm xúc. Ngoài ra, với mô hình biến đổi tiếng nói tiếng Việt, luật biến đổi được sử dụng chung cho các loại câu khác nhau, điều này có thể làm giảm tính tự nhiên của tiếng nói tổng hợp. Trong thời gian tới, chúng tôi sẽ tập trung giải quyết các hạn chế vừa nêu.

DANH MỤC CÁC CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ LIÊN QUAN ĐẾN LUẬN ÁN

1. Thi Duyen Ngo, The Duy Bui, (2009), **When and how to smile: Emotional expression for 3D conversational agents.** *Agent Computing and Multi-Agent Systems, volume 5044 of Lecture Notes in Computer Science,* chapter 31, pages 349-358. Springer Berlin/Heidelberg, Berlin, Heidelberg.
2. Thi Duyen Ngo, Nguyen Le Tran, Quoc Khanh Le, Chinh Huu Pham, Le Hung Bui, (2011), **An approach for building a Vietnamese talking face**. *Journal on Information and Communication Technologies*, ISSN 1859-3526, 6(26), pp. 207–216.
3. Thi Duyen Ngo, The Duy Bui, (2012), **A study on prosody of Vietnamese emotional speech.** *In Proceedings of the Fourth International Conference on Knowledge and Systems Engineering (KSE 2012)*, IEEE, pp. 151-155.
4. Thi Duyen Ngo, Masato Akagi, The Duy Bui, (2014), **Toward a Rule-Based Synthesis of Vietnamese Emotional Speech.** *In Proceedings of the Sixth International Conference on Knowledge and Systems Engineering (KSE 2014)*, Advances in Intelligent Systems and Computing 326, pp. 129-142, Springer International Publishing.
5. Thi Duyen Ngo, Thi Chau Ma, The Duy Bui. (2014), **Emotional facial expression analysis in the time domain.** *In Proceedings of the Sixth International Conference on Knowledge and Systems Engineering (KSE 2014)*, Advances in Intelligent Systems and Computing 326, pp. 487-498, Springer International Publishing.
6. Thi Duyen Ngo, Thi Hong Nhan Vu, Viet Ha Nguyen, The Duy Bui. (2014), **Improving simulation of continuous emotional facial expressions by analyzing videos of human facial activities.** *In Proc. of the 17th International Conference on Principles and Practice of Multi-Agent Systems (PRIMA 2014). Lecture Notes in Computer Science Volume 8861, 2014,* pp. 222-237. Springer International Publishing.
7. Thi Duyen Ngo, The Duy Bui. (2015), **A Vietnamese 3D Taking Face for Embodied Conversational Agents**. *In Proc. of the 11th IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF 2015),* pp.94-99.