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

LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

HÀ NỘI – 2015

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

LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC
PGS.TS. Bùi Thế Duy
GS.TS. Masato Akagi

HÀ NỘI – 2015

LỜI CẢM ƠN

Luận án được thực hiện tại Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội, dưới sự hướng dẫn của PGS.TS. Bùi Thế Duy và GS.TS. Masato Akagi.

Tôi xin gửi lời cảm ơn chân thành và sâu sắc nhất tới PGS. TS. Bùi Thế Duy – Bộ Khoa học và Công nghệ và GS. TS. Masato Akagi – Viện Khoa học và Công nghệ tiên tiến Nhật Bản (JAIST), những người thầy tâm huyết đã tận tình hướng dẫn, động viên khích lệ, dành nhiều thời gian quí báu để định hướng cho tôi trong quá trình tham gia khóa học và hoàn thiện luận án.

Tôi xin gửi lời cảm ơn chân thành tới lãnh đạo trường Đại học Công nghệ, lãnh đạo Khoa Công nghệ thông tin, cảm ơn các đồng nghiệp đã tạo điều kiện thuận lợi cho tôi trong quá trình làm luận án.

Tôi xin gửi lời cảm ơn chân thành tới các bạn đồng nghiệp trong phòng thí nghiệm Tương tác Người máy, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội, những người luôn bên tôi động viên, góp ý, chỉnh sửa trong quá trình viết luận án.

Cuối cùng, tôi xin gửi lời cảm ơn sâu sắc tới gia đình và bạn bè, những người đã luôn ủng hộ và hỗ trợ tôi về mọi mặt để tôi yên tâm học tập, nghiên cứu, và hoàn thành luận án.

LỜI CAM ĐOAN

Tôi xin cam đoan: Bản luận án tốt nghiệp này là công trình nghiên cứu thực sự của cá nhân. Các kết quả được viết chung với các tác giả khác đều được sự đồng ý của các đồng tác giả trước khi đưa vào luận án. Các kết quả nêu trong luận án là trung thực và chưa từng được công bố dưới bất cứ hình thức nào trước khi trình, bảo vệ và công nhận bởi "Hội Đồng đánh giá luận án tốt nghiệp Tiến sĩ Công nghệ Thông Tin".

Một lần nữa, tôi xin khẳng định về sự trung thực của lời cam kết trên.

Tác giả:

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT

VIẾT TẮT:

EFE (Emotional Facial Expression): Biểu cảm khuôn mặt thể hiện cảm xúc.

ES (Emotional State): Trạng thái cảm xúc.

ESV (Emotional State Vector): Véc tơ trạng thái cảm xúc.

FMCV (Facial Muscle Contraction Level): Véc tơ mức co cơ mặt.

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

AU (Action Unit): Đơn vị cử động.

3D (Three Dimensions) Ba chiều.

THUẬT NGỮ:

Nhân vật ảo: Nhân vật hoạt hình trong các ứng dụng giải trí, giáo dục, thương mại...

Embodied agent: Nhân vật ảo được thể hiện dưới hình dáng con người hoặc động vật.

Biểu cảm khuôn mặt: Một trạng thái khuôn mặt thể hiện cảm xúc nào đó.

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ

TÓM TẮT LUẬN ÁN

Luận án nghiên cứu những vấn đề xung quanh bài toán xây dựng nhân vật ảo. Cụ thể luận án tập trung giải quyết bài toán thể hiện cảm xúc cho nhân vật ảo nói tiếng Việt. Nhân vật ảo là kết quả của sự kết hợp giữa các lĩnh vực nghiên cứu như đồ họa máy tính, tác nhân tự động, công nghệ tiếng nói và ngôn ngữ. Các nhân vật ảo có khả năng giao tiếp này ngày càng phổ biến trong truyền thông đa phương tiện. Nhiều kỹ thuật đã và đang được phát triển nhằm tạo cho các nhân vật này có khả năng hành xử theo lối giống với con người. Để có thể đạt được điều đó, nhân vật ảo được mô phỏng với cảm xúc và cá tính, cũng như các kênh giao tiếp khác như tiếng nói, thao tác và biểu cảm khuôn mặt,... Để tăng tính thuyết phục, nhân vật ảo cần được cung cấp khả năng thể hiện cảm xúc. Tổng hợp các nghiên cứu cũng như thực tế cho thấy khuôn mặt và tiếng nói là hai kênh quan trọng nhất trong việc thể hiện cảm xúc của con người. Vì vậy, luận án tập trung vào hai kênh này khi 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 liên quan đến bài toán thể hiện cảm xúc cho nhân vật ảo nói tiếng Việt, 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.

Thứ hai, luận án đề xuất mô hình biến đổi tiếng nói tiếng Việt ở trạng thái không cảm xúc thành tiếng nói có cảm xúc, cung cấp cho nhân vật ảo nói tiếng Việt khả năng thể hiện cảm xúc trong kênh tiếng nói.

Thứ ba, luận án xây dựng một khuôn mặt ba chiều nói tiếng Việt cho nhân vật ảo. Việc này giúp cho nhân vật ảo có khả năng thể hiện trạng thái cảm xúc liên tục một cách tự nhiên qua biểu cảm khuôn mặt, cũng như có khả năng thể hiện cảm xúc trong giọng nói tiếng Việt.

Chương 1

Giới thiệu

1.1 Đặt vấn đề

Chúng ta thường xuyên xem những bộ phim do con người đóng, tuy nhiên thật khó để có thể tìm được một diễn viên có ngoại hình như nhân vật Shrek trong bộ phim hoạt hình *Shrek* mà lại có khả năng mang về doanh thu cao lên đến hàng trăm triệu đô la như loạt phim hoạt hình này. Điều gì khiến cho gã chằn tinh xấu xí và đáng sợ như Shrek có thể giành được nhiều tình cảm từ khán giả đến vậy? Có lẽ một trong những yếu tố quan trong nhất chính là chúng ta cảm thấy đồng cảm với Shrek. Nhìn vào Hình 1.1, thật khó để không có cảm tình với anh chàng chằn tinh xấu xí nhưng tốt bụng này. Cùng với sự thành công của một số bộ phim hoạt hình khác như *Gia Đình Nhà Siêu Nhân (The Incredibles)*, *Robot biết yêu (Wall-e)*, lĩnh vực hoạt hình mà trung tâm là việc tạo ra các nhân vật hoạt hình đã và đang nhận được sự quan tâm lớn.

Hình 1.1: Hai vợ chồng nhà "chằn tinh" Shrek.

Cũng liên quan đến các nhân vật ảo, nhưng không phải nhân vật hoạt hình mà là các nhân vật ảo trong máy tính. 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 dành nhiều công sức hơn nhằm cải tiến 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,... Khả năng về ngôn ngữ, biểu cảm khuôn mặt và cử chỉ của nhân vật ảo khiến cho chúng phù hợp với các ứng dụng này. Ví dụ, thế giới của các trò chơi nhập vai đang phát triển hơn lúc nào hết khi người chơi bật máy tính lên là có thể giao tiếp với các nhân vật ảo mà cảm giác như đang sống trong thế giới thực (Second Life, The Sims, Fallout 3). Nhân vật ảo cũng có thể được sử dụng trong ứng dụng giải trí với vai trò người kể chuyện ảo [140]. Ngoài ra, nhân vật ảo còn được sử dụng trong các ứng dụng giáo dục. Với ứng dụng mô phỏng phòng học ảo, nhân vật ảo có thể được sử dụng với vai trò người thầy để thực hiện các thao tác minh họa, trả lời các câu hỏi, và điều khiển việc học của các sinh viên [121]. Chúng cũng có thể được dùng trong vai trò bạn học để thực hiện các thao tác yêu cầu nhiều người. Nhân vật ảo còn có thể được dùng trong các ứng dụng thương mại điện tử, dịch vụ du lịch, hệ thống truy vấn dịch vụ... Vì những ứng dụng thực tế của mình, nhân vật ảo nhận được rất nhiều sự quan tâm, nghiên cứu.

Nhân vật hoạt hình là nhân vật được lập trình sẵn, những hành động, biểu cảm của chúng được tạo bởi các thao tác bằng tay của con người. Còn với nhân vật ảo trong máy tính, những hành động, phản ứng, biểu cảm... của chúng là do máy tính sinh ra một cách tự động. Luận án dùng thuật ngữ "nhân vật ảo" để chỉ các nhân vật ảo trong máy tính. Mục tiêu chung khi nghiên cứu về nhân vật ảo là khiến cho chúng trở nên thuyết phục hơn, theo cách làm cho hoạt động và phản ứng của chúng đối với người dùng là giống như trong thế giới thực. Nhiều kỹ thuật đã và đang được phát triển nhằm tạo cho các nhân vật ảo này này có khả năng hành xử theo lối giống với con người. Để có thể đạt được điều đó, nhân vật ảo được mô phỏng với các kênh giao tiếp như tiếng nói, cử động của đầu và mắt, các thao tác và biểu cảm khuôn mặt [5, 16, 29, 79]. Hơn nữa, ngoài chức năng nhận thức, chúng cũng được mô phỏng với cảm xúc và cá tính

[19, 29, 41, 119].

Chúng ta đã nói nhiều về nhân vật ảo, vậy nhân vật ảo là gì? Trước tiên, hãy xem qua khái niệm về "tác nhân" (agent) được định nghĩa bởi Wooldridge [153]:

> "Một tác nhân là một hệ thống máy tính được đặt trong một số môi trường, có khả năng hoạt động tự động trong môi trường đó để đáp ứng những mục tiêu đã được thiết kế."

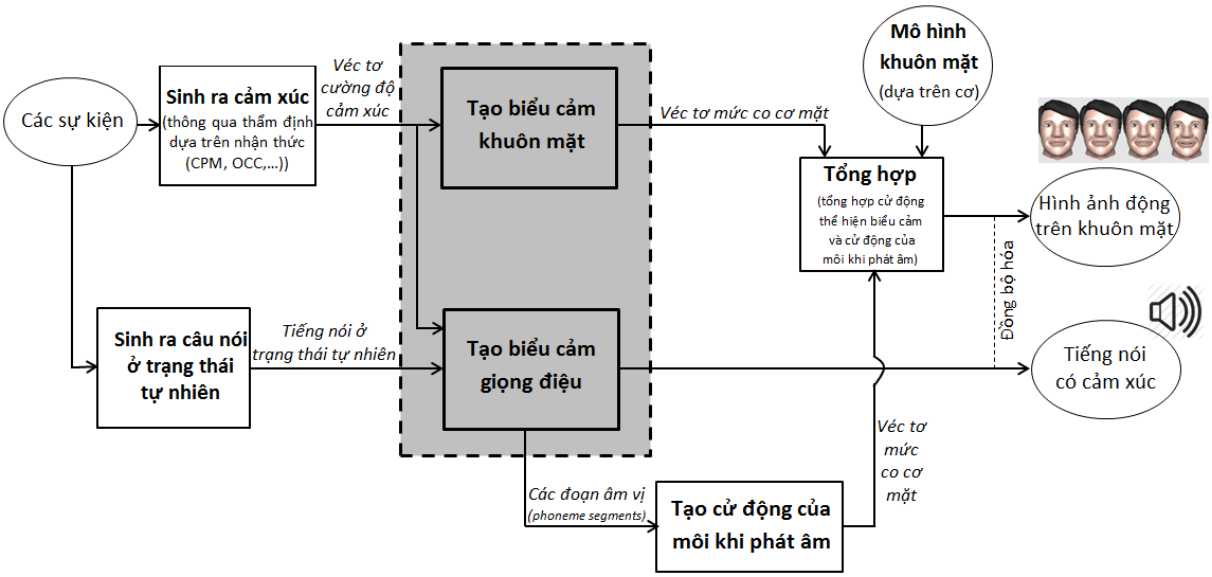
Nhân vật ảo là một loại tác nhân đặc biệt, nó được thể hiện dưới dạng cơ thể người hoặc cơ thể động vật được hoạt hóa, hay đôi khi chỉ là khuôn mặt có khả năng nói. Để xây dựng một nhân vật ảo, thông thường chúng ta cần xây dựng ba thành phần sau:

- 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.
- Một cơ thể có khả năng thể hiện những cử chỉ.
- Một mô hình trí tuệ bao gồm suy nghĩ, cảm xúc, động lực, hành vi, tính cách... của nhân vật.

Với bài toán xây dựng khuôn mặt và cơ thể thì khuôn mặt luôn nhận được nhiều sự quan tâm hơn vì khuôn mặt là nơi giao tiếp, nói chuyện, và bộc lộ cảm xúc. Khi quan sát nhân vật ảo, chúng ta thường quan sát khuôn mặt của những nhân vật đó nhiều hơn là quan sát cơ thể của chúng. 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

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. Trong mô hình này, cảm xúc của nhân vật ảo được thể hiện qua hai kênh chính nhất đó là khuôn mặt và tiếng nói. Dựa trên quá trình thẩm định các sự kiện đầu vào, mô đun "Sinh ra cảm



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

xúc" có chức năng *cung cấp trạng thái cảm xúc* cho nhân vật ảo. Từ đó, các mô đun còn lại *cung cấp cơ chế thể hiện cảm xúc* và tạo các biểu cảm thể hiện cảm xúc trên khuôn mặt và trong giọng nói cho nhân vật ảo. Nội dung nghiên cứu của luận án liên quan đến bài toán thể hiện cảm xúc cho nhân vật ảo, 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 mà luận án giải quyết 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, biểu cảm này được thể hiện trên khuôn mặt và trong giọng nói tiếng Việt. Đã có những nghiên cứu được đề xuất để giải quyết bài toán này. Hầu hết các nghiên cứu tập trung vào hai kênh biểu cảm chính đó là khuôn mặt và tiếng nói. Lý do là vì qua thực tế cũng như tổng hợp các nghiên cứu cho thấy đây là hai kênh quan trọng nhất trong việc thể hiện trạng thái cảm xúc. Tuy nhiên, đa số các nghiên cứu chỉ tập trung vào một kênh biểu cảm đơn thay vì quan tâm đến hai hay nhiều kênh biểu cảm khác nhau. 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.

Với kênh khuôn mặt, các nghiên cứu đã chỉ ra rằng biểu cảm khuôn mặt cho các cảm xúc cơ bản là phổ biến, có tính chất tương đồng giữa các nền văn hóa [36, 70]. Đến nay, nhiều nghiên cứu về thể hiện cảm xúc trên khuôn mặt

cho nhân vật ảo đã được đề xuất. Những nghiên cứu này có thể được chia thành hai lớp: phương pháp thể hiện cảm xúc tĩnh, và phương pháp thể hiện cảm xúc động. Phương pháp thể hiện cảm xúc tĩnh [4, 81, 83, 118] không có khả năng thể hiện trạng thái cảm xúc liên tục, không cung cấp một cơ chế nhất quán nào cho việc tạo biểu cảm trên khuôn mặt. Phương pháp thể hiện cảm xúc động [18, 80, 95, 119, 138, 147, 156] 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 thể hiện cảm xúc trên khuôn mặt, và giải quyết được hạn chế của phương pháp thể hiện cảm xúc tĩnh. Tuy nhiên, trong phương pháp này, biểu cảm khuôn mặt được tạo ra từ trạng thái cảm xúc liên tục theo cơ chế ánh xạ trực tiếp. Trong mỗi khoảng nhỏ thời gian, trạng thái cảm xúc được ánh xạ trực tiếp thành biểu cảm, sau đó biểu cảm này được thể hiện trên khuôn mặt. Cơ chế này sẽ tạo ra biểu cảm không tự nhiên khi có một trạng thái cảm xúc với cường độ cao xảy ra trong thời gian dài. Trong tình huống đó, biểu cảm có thể sẽ xuất hiện trên khuôn mặt trong thời gian khá dài; điều này có thể làm giảm tính tự nhiên của nhân vật ảo.

Với kênh tiếng nói, mục tiêu của bài toán là cung cấp cho nhân vật ảo khả năng tạo biểu cảm trong giọng nói tiếng Việt. Đã có một số nghiên cứu về ngôn điệu và âm sắc của tiếng nói tiếng Việt được đề xuất [65, 87, 88, 89, 93, 146]; một số nghiên cứu về tổng hợp tiếng nói tiếng Việt cũng được công bố [105, 151, 150]. Tuy nhiên, hầu hết các nghiên cứu này tập trung vào tiếng nói tiếng Việt ở trạng thái không cảm xúc. Theo hiểu biết của chúng tôi, đến nay chưa có nghiên cứu nào cung cấp khả năng thể hiện cảm xúc trong giọng nói tiếng Việt cho nhân vật ảo, và cũng chưa có nghiên cứu nào về xây dựng khuôn mặt cho nhân vật ảo nói tiếng Việt có khả năng thể hiện cảm xúc trên khuôn mặt và trong tiếng nói.

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

- Thứ nhất, để tăng tính tự nhiên, thuyết phục của biểu cảm khuôn mặt thể hiện cảm xúc cho nhân vật ảo, hạn chế nhược điểm của cơ chế ánh xạ trực tiếp nói trên, luận án đề xuất mô hình chuyển trạng thái cảm xúc liên tục của nhân vật ảo thành biểu cảm khuôn mặt. Mô hình đề xuất 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 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. Ý tưởng này xuất phát từ kết quả của 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. Quá trình phân tích cơ sở dữ liệu và mô hình đề xuất sẽ được trình bày cụ thể ở Chương 3. Mô hình chuyển trạng thái cảm xúc liên tục thành biểu cảm khuôn mặt không chỉ được áp dụng riêng cho nhân vật ảo có khuôn mặt người Việt mà còn có thể được sử dụng cho các nhân vật ảo với khuôn mặt của người châu Âu, châu Á nói chung, châu Mĩ,...

- Thứ hai, để cung cấp khả năng thể hiện cảm xúc trong kênh tiếng nói cho nhân vật ảo, luận án đề xuất mô hình biến đổi tiếng nói tiếng Việt ở trạng thái không cảm xúc thành tiếng nói có cảm xúc. Từ việc phân tích cơ sở dữ liệu tiếng nói tiếng Việt có cảm xúc, các hệ số thể hiện quan hệ giữa đặc trưng âm của trạng thái không cảm xúc và đặc trưng âm của trạng thái cảm xúc được đưa ra. Từ đó, tập các luật dùng để chuyển tiếng nói không cảm xúc thành tiếng nói có cảm xúc được xây dựng. Từ tập các luật này, kỹ thuật biến đổi tiếng nói được sử dụng để tổng hợp tiếng nói tiếng Việt có cảm xúc từ tiếng nói ở trạng thái không cảm xúc. Quá trình phân tích cơ sở dữ liệu tiếng nói và tổng hợp tiếng nói có cảm xúc sẽ được trình bày ở Chương 4. Mô hình biến đổi tiếng nói tiếng Việt không cảm xúc thành tiếng nói có cảm xúc được sử dụng tạo biểu cảm trong giọng nói cho các nhân vật ảo nói tiếng Việt.

- 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. Ngoài việc tích hợp kết quả nghiên cứu từ Chương 3 và Chương 4, luận án cũng xây dựng hệ thống hình vị tiếng Việt để cung cấp cho nhân vật ảo khả năng thể hiện cử động của môi khi phát âm các từ tiếng Việt. 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. Quá trình xây dựng và đánh giá khuôn mặt ba chiều này được trình bày chi tiết trong Chương 5.

1.3 Cấu trúc của luận án

Phần còn lại của 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. Đây là cơ sở lý thuyết cho cho việc xây dựng các mô hình thể hiện cảm xúc trên khuôn mặt và trong tiếng nói sẽ được trình bày ở Chương 3 và Chương 4 của luận án. Trong chương này, chúng tôi cũng tổng kết các nghiên cứu liên quan tới việc cung cấp cảm xúc cho nhân vật ảo.

Trong Chương 3, luận án trình bày 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 hình thứ hai. Trước tiên luận án mô tả quá trình phân tích một cơ sở dữ liệu video tự nhiên; cơ sở dữ liệu này gồm các file video thể hiện khuôn mặt người thật biểu cảm các trạng thái cảm xúc khác nhau. Từ kết quả phân tích, luận án đưa ra các "mẫu" biểu cảm theo thời gian của các cảm xúc cơ bản. Dựa trên các mẫu biểu cảm này, mô hình tạo biểu cảm khuôn mặt thể hiện trạng thái cảm xúc liên tục được đề xuất.

Trong Chương 4, luận án 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 không cảm xúc. Trước tiên, luận án mô tả quá trình phân tích một cơ sở dữ liệu tiếng nói tiếng Việt nhân tạo; cơ sở dữ liệu này gồm các file audio chứa phát âm tiếng Việt ở các trạng thái cảm xúc khác nhau. Từ việc phân tích cơ sở dữ liệu, các luật thể hiện mối quan hệ về đặc trưng âm giữa tiếng nói có cảm xúc và tiếng nói ở trạng thái không cảm xúc được xây dựng. Từ đó luận án đề xuất mô hình biến đổi phát âm tiếng Việt ở trạng thái không cảm xúc thành phát âm tiếng Việt có cảm xúc.

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 có khả năng thể hiện cảm xúc trong giọng nói tiếng Việt. Ngoài ra, để xây dựng khuôn mặt ba chiều, một hệ thống hình vị tiếng Việt cũng được tổng hợp để cung cấp cho nhân vật ảo khả năng thể hiện cử động của môi khi phát âm các từ tiếng Việt.

Chương 2

Cảm xúc và thể hiện cảm xúc cho nhân vật ảo

Một trong những đặc điểm đặc của con người là có cảm xúc, điều này khiến con người khác với các động vật khác. Cảm xúc đã được nghiên cứu trong một thời gian dài và các kết quả chỉ ra rằng chúng đóng vai trò quan trọng trong chức năng nhận thức của con người. Cảm xúc mạnh tới mức chúng có thể ảnh hưởng tới tính sáng tạo, sự đánh giá, việc đưa ra quyết định, giao tiếp, và các tiến trình nhận thức khác của con người [27, 48, 52]. Chúng có thể đẩy con người tới hành động dũng cảm hoặc cực kỳ thô bạo và hướng hành động của con người theo cách này hay cách khác. Điều này đã được chỉ ra trong "Affective computing" của Picard [115].

Trong thực tế, cảm xúc có một vai trò cực kỳ quan trọng trong suốt quá trình giao tiếp của con người. Một điều khá rõ ràng rằng hiểu được con người diễn tả cảm xúc như thế nào và những biểu cảm này có ý nghĩa gì rất quan trọng, đôi khi mang tính chất cốt yếu trong giao tiếp xã hội thông thường. Hầu hết mọi người đều đặc biệt quan tâm tới việc người khác nghĩ gì và cảm thấy thế nào về họ, và họ theo dõi hành vi của người khác một cách cẩn thận. Có nhiều mối quan hệ bạn bè phát triển từ việc cảm nhận được cảm xúc tích cực. Cũng có nhiều mối quan hệ khác rơi vào tình trạng căng thẳng, đổ vỡ, hay thậm chí là chấm dứt vì những hiểu nhầm trong hành động và cử chỉ biểu cảm. Cảm xúc có thể không chỉ ra được hành động cụ thể mà một người sắp thực hiện nhưng nó có thể cho biết loại hành động mà người đó có thể thực hiện. Chúng thể hiện bối cảnh bên trong một con người mà ở đó những suy nghĩ của con người được xử lý. Về khía cạnh tiến hóa, việc phát hiện cảm xúc từ một cái nhìn riêng biệt có thể không có ý nghĩa, nhưng vì là một nhóm mang tính xã hội, con người có thể tận dụng việc thể hiện cảm xúc để xây dựng và duy trì các mối quan hệ.

Mặc dù hiểu cách mà cảm xúc được diễn tả không giúp chúng ta đọc suy nghĩ nhưng nó sẽ giúp chúng ta có được cách tốt nhất để sử dụng nguồn thông tin tức thời duy nhất chúng ta có – hành vi biểu cảm. Mọi người thường xem xét trạng thái cảm xúc của người khác, có thể bởi vì đó là dấu hiệu tốt chỉ ra hiện tại người đó cảm thấy thế nào, tiếp theo họ có thể làm gì, và họ có thể làm như thế nào. Với đánh giá này, khuôn mặt con người trở thành phần có tính giao tiếp nhất trên cơ thể trong việc thể hiện cảm xúc, cung cấp nhiều thông tin hỗ trợ hội thoại [36]. Các kết quả nghiên cứu đã chỉ ra rằng cử động khuôn mặt đóng vai trò quan trọng trong thể hiện cảm xúc. Người ta nhận thấy rằng tồn tại một mối liên kết giữa cử động khuôn mặt và trạng thái cảm xúc; điều này được khẳng định trong bài viết của Darwin [28]. Ngoài ra, các nghiên cứu cũng chỉ ra rằng tồn tại mối liên kết giữa đặc trưng âm của giọng nói và trạng thái cảm xúc [24]. Đây cũng là một trong những kênh biểu cảm quan trọng trong việc thể hiện cảm xúc của con người.

Chương này của luận án tổng kết các nghiên cứu liên quan đến cảm xúc và mối quan hệ giữa trạng thái cảm xúc với các kênh biểu cảm. Đây là cơ sở lý thuyết quan trọng cho việc xây dựng mô hình cảm xúc cũng như mô hình tạo biểu cảm thể hiện cảm xúc cho nhân vật ảo. Ngoài ra, các nghiên cứu liên quan tới bài toán cung cấp cảm xúc cho nhân vật ảo cũng được tổng kết. Nội dung của chương được tổ chức như sau. Phần 2.1 trình bày tổng quan các nghiên cứu tâm lý về cảm xúc. Sau đó, Phần 2.2 tóm tắt các nghiên cứu tâm lý và sinh lý học về mối quan hệ giữa trạng thái cảm xúc với biểu cảm thể hiện cảm xúc. Tiếp theo, Phần 2.3 trình bày tổng quan các nghiên cứu cung cấp cảm xúc cho nhân vật ảo. Cuối cùng, Phần 2.4 nêu kết luận chương và bàn luận.

2.1 Nghiên cứu tâm lý học về cảm xúc

Theo các nghiên cứu tâm lý học, có khá nhiều quan điểm khác nhau trong cách nhìn nhận về cảm xúc. Trong [26], Cornelius đã tổng kết bốn quan điểm chính về mặt tâm lý học để định nghĩa, nghiên cứu, và giải thích về cảm xúc. Bốn quan điểm này là: *học thuyết Darwin (Darwinian)*, *học thuyết Henry James (Jamesian)*, *quan điểm kiến tạo xã hội (social constructivist)*, và *quan điểm nhận thức (cognitive)* về cảm xúc. Những quan điểm này đều có tập giả định riêng để

thể hiện các cách suy nghĩ khác nhau về cảm xúc.

Quan điểm Darwin được đề xướng bởi Darwin [28] cho rằng cảm xúc là phổ quát và có các chức năng thích nghi. Quan điểm này tập trung vào chức năng của cảm xúc trong bối cảnh tiến hóa bởi chọn lọc tự nhiên. Darwin quả quyết rằng các cơ chế hành xử mà chúng ta vẫn xem là "biểu cảm thể hiện cảm xúc" tiến hóa không phải cho mục đích thể hiện cảm xúc mà là cho các mục đích khác; chúng được xem là "cảm xúc" bởi vì chúng xảy ra cùng với các hành động gắn liền với các cảm xúc mạnh. Ví dụ cách thức biểu hiện sự ghê tởm của con người trên khuôn mặt cũng tương tự như biểu hiện trên khuôn mặt của một con chó khi nó ngửi hay nếm phải thức ăn mà nó không thích. Lý do của sự tương tự này là cả hai khuôn mặt đều gắn với hành động là tống thức ăn đó ra. Các nghiên cứu chỉ ra rằng những người đến từ các nền văn hóa khác nhau có thể nhận diện các biểu cảm khuôn mặt của một số lượng nhỏ các cảm xúc [31, 73]. Những nghiên cứu này nhấn mạnh rằng tính phổ quát của cảm xúc là một phần trong tiến hóa của con người. Khi giận dữ, khuôn mặt mà một người nào đó tạo ra sẽ giống với khuôn mặt do những người khác tạo ra bởi vì khuôn mặt như vậy là công cụ giao tiếp quan trọng trong suốt lịch sử loài của chúng ta. Một số nhà nghiên cứu cũng xem xét chức năng thích nghi của cảm xúc. Trong [117], Plutchik chỉ ra tầm quan trọng của hành vi cảm xúc trong quá trình chọn lọc tự nhiên của tất cả các loài; cảm xúc được xem như sự thích nghi với các sự kiện trong cuộc sống. Cùng quan điểm với Darwin, thuyết "khuynh hướng hành động" của Frijda [51] xem cảm xúc là sự nhận biết khuynh hướng hành động. Khuynh hướng hành động gắn kết chặt chẽ với cách mà một người cảm nhận hay thẩm định môi trường. Thuyết tiến hóa của cảm xúc được đề xuất bởi Shaver và cộng sự [132] bắt đầu với giả định tương tự như giả định của Plutchik rằng tất cả mọi người đều có chung một tập các phản ứng nguyên mẫu với môi trường. Đi theo quan điểm của Frijda rằng cảm xúc là "khuynh hướng hành động" theo sau quá trình thẩm định môi trường của một người, Shaver và các cộng sự cho rằng có một tập nhỏ các cảm xúc cơ bản được nhận diện bởi tất cả các nền văn hóa.

Quan điểm James được đề xướng bởi James [74] xem cảm xúc như là các phản ứng của cơ thể; James 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 (như khóc, cười), hành vi công cụ (như chạy trốn hay thu mình lại), và những thay đổi sinh lý học (như run rẩy). Đi theo quan điểm này, các phương pháp tiếp cận hiện đại coi những thay đổi "bản năng" và hành vi biểu cảm là những thay đổi của cơ thể. Thay đổi "bản năng" là sự kích thích trong hệ thống thần kinh giao cảm - một nhánh của hệ thần kinh tự trị(autonomic nervous system - ANS). Những thay đổi này thể hiện các hành động và ảnh hưởng của các hành động đó lên tim, dạ dày, và các cơ quan khác bị chi phối bởi hệ thần kinh giao cảm [55, 133]. Hành vi biểu cảm là những thay đổi trong điệu bộ và biểu cảm khuôn mặt [72, 82]. Từ *quan điểm Jamesian*, một số nghiên cứu đã chứng minh rằng một tập nhỏ các cảm xúc như sợ, giận, buồn, vui có thể được phân biệt với nhau nhờ các mẫu cử động tự trị [39, 91]. Các nghiên cứu này cũng cho rằng phản ứng của hệ thần kinh tự trị giúp xác định cường độ của cảm xúc được trải nghiệm. Allport [6], Izard [71] và các nhà nghiên cứu khác chỉ ra rằng phản ứng từ khuôn mặt cũng có thể được sử dụng để xác định cường độ của cảm xúc và để phân biệt các cảm xúc.

Quan điểm kiến tạo xã hội được đề xuất đầu tiên bởi Averill [9] xem xét cảm xúc như các thành phần xây dựng xã hội, phục vụ cho các mục đích xã hội; cảm xúc được xem như "một vai trò xã hội tạm thời bao gồm đánh giá, thẩm định của cá nhân về tình huống, và vai trò này được xem như cảm xúc chứ không phải hành động". Trái với các giả định của *quan điểm Darwinian* và *quan điểm Jamesian* cho rằng cảm xúc chủ yếu là các hiện tượng sinh học, *quan điểm kiến tạo xã hội* 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. Ví dụ, nếu một người bị xúc phạm bởi một trong số bạn tốt của anh ta khi có sự hiện diện của vài người khác, thường thì anh ta sẽ trở nên giận dữ nếu anh ta trưởng thành ở Mỹ hay một quốc gia phương Tây khác; tuy nhiên, nếu anh ta trưởng thành ở Nhật thì có thể anh ta chỉ đơn giản là mỉm cười người bạn khiến anh ta tức giận. Theo Cornelius [26], điểm chính của *quan điểm kiến tạo xã hội* đó là trải nghiệm và thể hiện cảm xúc phụ thuộc vào các qui ước hay qui tắc được học; những qui ước, qui tắc này là khác nhau ở các nền văn hóa. Các nghiên cứu theo *quan điểm kiến tạo xã hội* cho rằng biểu cảm thể hiện cảm xúc có sự biến đổi giữa các nền văn hóa. Vấn đề này thường bị chất vấn bởi các nhà nghiên cứu theo *quan điểm Darwinian* và *quan điểm Jamesian* - những người tin rằng có sự phổ quát nhất định trong biểu cảm thể hiện cảm xúc. Tuy nhiên, bằng chứng

của sự biến đổi hay sự phổ quát vẫn còn đang được tranh luận.

Quan điểm nhận thức được đưa ra trước tiên bởi Arnold [8] 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. Arnold [8] cho rằng cảm xúc thực ra được bắt đầu bởi sự đánh giá của một người về hoàn cảnh/tình huống của anh ta/cô ta. Arnold nhận thấy trải nghiệm trong quá khứ và mục tiêu của một người là những nhân tố quan trọng trong cách mà người đó đánh giá một tình huống. Ngay sau đó, Speisman và cộng sự [136] đã thực hiện một loạt nghiên cứu đi theo quan điểm của Arnold đó là đặc trưng phản ứng cảm xúc của một người đối với một sự kiện phụ thuộc vào việc người đó đánh giá sự kiện như thế nào. Các kết quả nghiên cứu này sau đó được nhân rộng và mở rộng bởi Lazarus và Alfert [85, 86, 84]. Ý tưởng chính trong học thuyết cảm xúc của Lazarus đó là cảm xúc là sau nhận thức (post-cognitive). Cho tới nay, có nhiều nhà nghiên cứu đi theo *quan điểm nhận thức*, ví dụ như Mandler[97], Oatley và Johnson-Laird [108],... Trong số các thuyết nhận thức về cảm xúc, có nhiều nghiên cứu [51, 84, 110, 122] liên quan tới việc đặc tả một cấu trúc nhận thức (cognitive structure) gắn với cảm xúc. Kết quả của những nghiên cứu này không chỉ có tầm quan trọng về mặt lý thuyết mà còn có ý nghĩa thực tế trong nhiều lĩnh vực khác nhau như liệu pháp tâm lý trong y học, và trí tuệ nhân tạo.

Bốn quan điểm nói trên giúp chúng ta hiểu hơn về bản chất của cảm xúc. Có sự trùng lặp ở mức độ nào đó giữa bốn quan điểm này, và không phải tất cả các nghiên cứu về cảm xúc chỉ đi theo một quan điểm duy nhất. Có một số nghiên cứu đi theo hai hay ba quan điểm. Ví dụ như nghiên cứu của Ekman [35] đi theo cả hai quan điểm là Darwinian và Jamesian để tìm hiểu về cảm xúc. Những quan điểm này là nền tảng cơ sở cho các nghiên cứu về mối quan hệ giữa cảm xúc và các kênh biểu cảm được tổng kết ở phần tiếp theo.

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

Diễn đạt cảm xúc thường được xem như là một khía cạnh của giao tiếp lời nói và phi lời nói. Con người có thể sử dụng từ ngữ để nói với người khác về cảm xúc của họ, nhưng họ cũng có thể truyền tải cảm xúc thông qua giọng điệu khi nói và thông qua các kênh phi lời nói như biểu cảm khuôn mặt, cử động cơ thể, và điệu bộ.

Một trong những vấn đề chính liên quan đến diễn đạt cảm xúc là có rất ít hành vi liên quan riêng biệt đến cảm xúc. Các hành vi thường có vô số ý nghĩa và khía cạnh cảm xúc chỉ là một trong số đó. Ví dụ như giao tiếp của mắt được sử dụng khi nói để điều hòa luồng hội thoại qua lại, nhưng nó cũng diễn tả thái độ giữa các cá nhân như chống đối hay thu hút. Ngôn ngữ được sử dụng để truyền tải các ý tưởng nhưng nó cũng bao hàm cả các đặc trưng diễn tả cảm nhận của người nói về thông điệp, hội thoại, và con người được đề cập đến. Diễn đạt trên khuôn mặt của một người, dáng điệu cơ thể, và cử chỉ cung cấp thông tin về các đặc tính tương đối tĩnh (như địa vị xã hội, nghề nghiệp, và cá tính) cũng như các trạng thái có tính tạm thời hơn như là cảm xúc và tâm trạng. Không chỉ diễn tả đồng thời nhiều thông điệp mà cùng một hành vi có thể có nhiều ý nghĩa khác nhau. Mặc dù diễn đạt cảm xúc chỉ là một trong nhiều khía cạnh của thông điệp, tuy vậy nó vô cùng phức tạp và điều này dẫn tới sự phân chia cần thiết trong việc nghiên cứu.

Ban đầu, hầu hết các nhà nghiên cứu có khuynh hướng tập trung vào một kênh biểu cảm tại một thời điểm và xem xét những kênh này một cách riêng biệt. Xu hướng này dường như đã thay đổi trong những năm gần đây và nhiều nghiên cứu hiện nay liên quan tới đóng góp của các kênh biểu cảm trong biểu cảm tổng thể [24]. Thuật ngữ "kênh" (channel) được sử dụng khá thường xuyên nhưng hiếm khi nó được cho một định nghĩa chính xác. Wiener & Mehrabian (1968) định nghĩa kênh là "một tập bất kỳ của các hành vi trong giao tiếp, được chỉ rõ một cách có hệ thống bởi một người quan sát, và được xem xét bởi người đó để truyền tải các thông tin có thể được nghiên cứu (ít nhất về mặt nguyên lý) độc lập với các hành vi khác xảy ra tại cùng thời điểm". Ý tưởng quan trọng của định nghĩa đó là *các kênh có thể được tách rời*(ít nhất là về mặt nguyên lý). Một ví dụ trong lĩnh vực nghiên cứu về diễn đạt cảm xúc là có thể phân biệt

giữa kênh liên quan đến thính giác và kênh liên quan đến thị giác. Các tín hiệu thính giác được tạo ra bằng cách dùng lời nói, được truyền tải bởi âm thanh, và đón nhận thông qua việc nghe; trong khi đó giao tiếp thị giác liên quan tới cử động khuôn mặt và cơ thể, được đón nhận bởi hành động nhìn. Các kênh thính giác có thể được chia nhỏ hơn thành đặc trưng phát âm, đặc trưng ngữ pháp, và nội dung; các kênh thị giác bao gồm biểu cảm khuôn mặt, cử động cơ thể, hành vi nhìn, và kiểm soát không gian cá nhân. Tập trung vào các kênh biểu cảm riêng biệt không có nghĩa là làm giảm tính phức tạp của các diễn đạt cảm xúc. Hầu hết các nhà nghiên cứu hiểu rằng các kênh diễn đạt cảm xúc là sự trừu tượng hóa cần thiết (sự trừu tượng hóa này không tồn tại trong giao tiếp xã hội thông thường). Tuy vậy, các kênh đóng vai trò là phương tiện hữu ích cho việc chia nhỏ nghiên cứu. Chúng ta (là các nhà khoa học) có thể lựa chọn tập trung vào một khía cạnh này hay khía cạnh kia, việc này thực ra là tự đặt giới hạn cho bản thân, được thêm vào luồng thông tin liên tục.

Thực tế cũng như các nghiên cứu cho thấy khuôn mặt và tiếng nói là hai kênh quan trọng nhất trong việc thể hiện cảm xúc [24]. Vì vậy, luận án tập trung và hai kênh này khi giải quyết bài toán thể hiện cảm xúc cho nhân vật ảo. Các nghiên cứu về mối quan hệ giữa trạng thái cảm xúc với hai kênh diễn đạt này sẽ được tổng kết sau đây.

2.2.1 Cảm xúc và cử động khuôn mặt

Cử động khuôn mặt đóng vai trò quan trọng trong giải thích nội dung hội thoại và thể hiện cảm xúc. Chúng xuất hiện một cách liên tục trong quá trình tương tác xã hội, đặc biệt là trong hội thoại. Chúng bao gồm cử động của môi khi nói, các tín hiệu giao tiếp, các diễn đạt cảm xúc, và các thao tác đáp ứng nhu cầu sinh học. Các kết quả nghiên cứu đã chỉ ra rằng tồn tại mối quan hệ giữa các cử động trên khuôn mặt với trạng thái cảm xúc của con người [28].

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*, *quan điểm nhận thức*, và *quan điểm đa chiều*.

Quan điểm cảm xúc cơ bản

Theo tổng kết của Kappas [76], các nhà nghiên cứu theo *quan điểm cảm xúc cơ bản* [36, 34, 70, 69, 144, 145] cho rằng có một tập nhỏ các cảm xúc có thể phân biệt hoàn toàn với nhau nhờ biểu cảm khuôn mặt. Ví dụ, khi một người vui thì anh ta cười, khi giận thì anh ta tỏ vẻ mặt khó chịu và không hài lòng. Russell và Fernández-Dols [124] đã tóm tắt *quan điểm cảm xúc cơ* bản như sau:

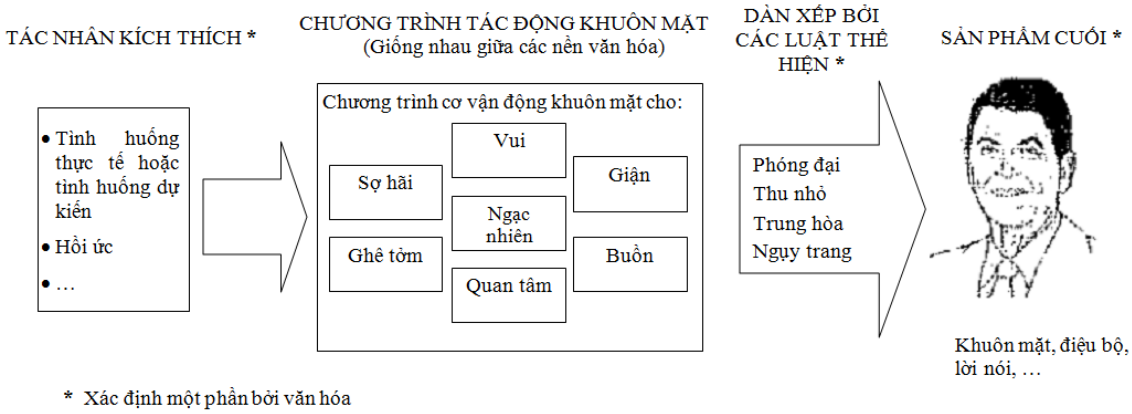
> "Mỗi cảm xúc cơ bản là hoàn toàn phổ quát và riêng biệt về mặt di truyền học. Mỗi cảm xúc là một mẫu rất chặt chẽ bao gồm hành vi khuôn mặt đặc trưng, sự từng trải có ý thức đặc biệt (cảm giác), nền tảng sinh lý học, và các hành động có ý nghĩa đặc trưng khác."

Một số điểm quan trọng của quan điểm này là:

- Tồn tại một mẫu biểu cảm nhất quán, bẩm sinh, và phổ quát cho mỗi cảm xúc cơ bản. Mẫu biểu cảm này sẽ xuất hiện khi cảm xúc cơ bản đó xuất hiện, trừ trường hợp ẩn hay che giấu cảm xúc. Vì vậy, trạng thái cảm xúc của một người có thể quan sát được từ khuôn mặt của người đó, trừ trường hợp ẩn hay che giấu cảm xúc.
- Trạng thái nào mà không có dấu hiệu khuôn mặt của riêng nó thì không phải là một cảm xúc cơ bản.
- 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.

Như vậy, theo *quan điểm cảm xúc cơ bản*, biểu cảm khuôn mặt của các cảm xúc cơ bản là nhất quán. Các biểu cảm này giống nhau ở tất cả mọi người, không phân biệt tuổi tác, trình độ, giới tính, nền văn hóa... Nghiên cứu của Ekman chỉ ra rằng có sáu biểu cảm khuôn mặt nhất quán, tương ứng với sáu cảm xúc cơ bản là vui, buồn, giận dữ, ngạc nhiên, ghê tởm, và sợ hãi [36].

Lấy *quan điểm cảm xúc cơ bản* làm trọng điểm, Ekman [31, 32, 33] đã đề xuất một mô hình *neurocultural* để giải thích tính cộng đồng và sự biến đổi trong biểu cảm khuôn mặt của con người, giải thích quan hệ giữa cảm xúc và biểu cảm khuôn mặt. Mô hình này được mô tả bởi Fridlund [49] (Hình 2.1) như sau:



Hình 2.1: Quan điểm của Ekman về quan hệ giữa cảm xúc và biểu cảm (minh họa bởi Fridlund [49]).

Trong mô hình này, khuôn mặt hàng ngày là kết quả của các biểu cảm khuôn mặt bẩm sinh, nhất quán thể hiện trạng thái cảm xúc, nhưng khuôn mặt ấy có thể bị thay đổi bởi yếu tố xã hội. Nói cách khác, trong mô hình này, thành phần *chương trình tác động khuôn mặt* thiết lập sự tương ứng phổ quát và bẩm sinh giữa trạng thái cảm xúc và mẫu khuôn mặt. Các cảm xúc pha trộn có thể xuất hiện, dẫn tới sự pha trộn của các mẫu khuôn mặt. Trước khi được thể hiện trên khuôn mặt, các mẫu khuôn mặt có thể được điều chỉnh bởi thành phần *dàn xếp bởi các luật thể hiện* của mô hình. Thành phần này có thể chặn hoặc làm yếu các biểu cảm khuôn mặt của cảm xúc, hoặc nó có thể tạo ra một biểu cảm "mặt nạ" để làm mờ hay thậm chí là loại bỏ một trạng thái cảm xúc.

Mặc dù mô hình của Ekman thể hiện được trọng tâm của *quan điểm cảm xúc cơ bản*, mô tả được mối quan hệ giữa cảm xúc và biểu cảm, nhưng Fridlund đã nêu ra hai nhược điểm của mô hình này. Thứ nhất, trong mô hình này không có tiêu chuẩn để xác định khi nào thì thành phần *chương trình tác động khuôn mặt* được kích hoạt, khi nào thì sự pha trộn các cảm xúc được tạo ra, và khi nào thì thành phần *dàn xếp bởi các luật thể hiện* hoạt động. Thứ hai, mô hình tạo ra mối quan hệ mơ hồ giữa biểu cảm khuôn mặt và cảm xúc. Trong khi các cảm xúc cơ bản được phát hiện dựa trên việc tìm kiếm các khuôn mặt phổ quát thì thành phần *dàn xếp bởi các luật thể hiện* khiến cho sự phát hiện này trở nên mơ hồ, trừ phi có một tiêu chuẩn rõ ràng để xác định khi nào thì thành phần này không được sử dụng. Có lẽ, khi dựa trên *quan điểm cảm xúc cơ bản*

để mô phỏng biểu cảm khuôn mặt thể hiện cảm xúc cho các nhân vật ảo, các nhà nghiên cứu cần phải quan tâm đến giải pháp để giải quyết hai nhược điểm này.

Quan điểm cảm xúc cơ bản cũng bị phê phán nhiều từ những nhà nghiên cứu đi theo các quan điểm khác. Đặc biệt, Russell[124] đứng trên *quan điểm đa chiều* đã đưa ra một số lý do để phản đối *quan điểm cảm xúc cơ bản*, ví dụ như về khái niệm cảm xúc riêng biệt, tính phổ quát của biểu cảm thể hiện cảm xúc cơ bản,... Tuy nhiên, bản thân những lập luận này không phải là không thể nghi ngờ. Ví dụ như đối với khái niệm cảm xúc riêng biệt, Russell cho rằng ông chưa bao giờ nhìn thấy bất kỳ bằng chứng nào về nó; về việc những đứa trẻ có khả năng phân biệt được các mẫu cử động khuôn mặt khác nhau, Russell tin rằng đó là dựa trên niềm vui và sự hưng phấn chứ không phải dựa trên cảm xúc riêng biệt. Có lẽ tranh luận mạnh mẽ nhất, có tính thuyết phục nhất chính là sự quả quyết rằng nếu không có ngữ cảnh thì biểu cảm khuôn mặt sẽ mơ hồ, nhập nhằng. Tranh luận này nhận được sự ủng hộ của Fernández-Dols và Carroll [46]. Từ đó, Russell và Fernández-Dols khăng khăng rằng những phát hiện trong *quan điểm cảm xúc cơ bản* là chưa đủ để các nhà nghiên cứu có thể hiểu một cách đầy đủ về biểu cảm khuôn mặt thể hiện cảm xúc.

Quan điểm nhận thức

Quan điểm nhận thức về biểu cảm khuôn mặt thể hiện cảm xúc được đề xuất bởi các nhà nghiên cứu theo *quan điểm nhận thức* (cognitive perspective) khi nghiên cứu về cảm xúc, ví dụ như Arnold [8] và Scherer [127]. Như đã đề cập trong Phần 2.1, *quan điểm nhận thức* cho rằng cảm xúc được kích hoạt bởi quá trình đánh giá/thẩm định nhận thức của một tình huống cá nhân. Khác với giả định của *quan điểm cảm xúc cơ bản* cho rằng hành động khuôn mặt được tạo ra theo các mẫu do một cảm xúc xác định đã được kích hoạt, các nhà nghiên cứu theo thuyết thẩm định 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. Ví dụ, mẫu khuôn mặt cau mày được tạo ra khi có điều không mong muốn xảy ra khiến chúng ta không đạt được mục đích [128, 135].

Quan điểm đa chiều

Quan điểm đa chiều được đề xuất bởi các nhà nghiên cứu có niềm tin rằng các trạng thái cảm xúc về cơ bản được phân biệt dựa trên một số lượng nhỏ các chiều như độ hấp dẫn nội tại (valence) và độ kích hoạt (activation), và cho rằng cử động khuôn mặt được liên kết với những chiều này (ví dụ Russell [124]). Quan điểm đa chiều cho rằng các chiều cơ bản của của một trạng thái cảm xúc bên dưới được phản ánh trong hành vi khuôn mặt. Ví dụ, Russell [124] biện luận rằng trong một số tình huống, hành vi khuôn mặt thay đổi một cách đơn giản hướng tới trạng thái hài lòng hay không hài lòng chứ không phải do các cảm xúc riêng biệt như buồn và vui. Quan điểm này dường như gần với *quan điểm nhận thức* hơn.

Vẫn tồn tại những tranh cãi kéo dài giữa ba quan điểm về biểu cảm khuôn mặt thể hiện cảm xúc. Tuy nhiên, mặc dù mỗi quan điểm có sự dự đoán riêng của mình, không phải là không thể có những ý tưởng chung giữa ba quan điểm này. Theo Ortony và cộng sự [110], mối quan hệ giữa các thành phần thẩm định với các chiều cơ bản và với các cảm xúc riêng biệt có vẻ phức tạp nhưng lại rất hợp lý. Ví dụ, Arnold [8] đề xuất chiều độ hấp dẫn nội tại (valence dimension) trong *quan điểm nhận thức*, cái này rõ ràng là có liên quan tới chiều độ hấp dẫn nội tại cơ bản (basic valence dimension) có trong các cách tiếp cận đa chiều.

Dường như không có nhà nghiên cứu từ bất kỳ quan điểm nào có thể đưa ra các bằng chứng để bảo vệ một cách đầy đủ những giả thuyết của họ. Mặc dù vậy, các nghiên cứu tâm lý học từ những quan điểm này có ảnh hưởng đáng kể đối với sự hiểu biết của chúng ta về mối liên kết giữa trạng thái cảm xúc và cử động khuôn mặt. Những nghiên cứu này đóng vai trò rất quan trọng trong thao tác mô phỏng và nhận dạng biểu cảm khuôn mặt thể hiện cảm xúc trên máy tính. Theo Kappas [76], *quan điểm cảm xúc cơ bản* là hữu ích nhất trong bối cảnh dự đoán cảm xúc từ cử động trên khuôn mặt. So với các nghiên cứu thuộc *quan điểm nhận thức* và *quan điểm đa chiều*, nghiên cứu thuộc *quan điểm cảm xúc cơ bản* cung cấp nhiều bằng chứng thực nghiệm về mối quan hệ giữa cảm xúc và cử động khuôn mặt hơn. Hơn nữa, những dự đoán của *quan điểm cảm xúc cơ bản* thường rất rõ ràng để xác nhận hoặc từ chối. Trong khi đó, nhiều dự đoán của *quan điểm nhận thức* và *quan điểm đa chiều* là không đủ cụ thể. Theo quan điểm của chúng tôi, trong việc mô phỏng mối quan hệ giữa cảm xúc

và cử động khuôn mặt, các kết quả nghiên cứu thuộc *quan điểm cảm xúc cơ bản* là hữu ích nhất. Tuy nhiên, với việc sử dụng máy tính để nhận dạng cảm xúc thì *quan điểm cảm xúc cơ bản* có thể không phải là lựa chọn tốt nhất. Lý do là vì có những cử động khuôn mặt không liên quan tới cảm xúc, ví dụ như tiếng nói trực quan hay các tín hiệu hội thoại.

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

Để nắm bắt được một cách khách quan sự phong phú và phức tạp của biểu cảm khuôn mặt, các nhà nghiên cứu nhận thấy rằng cần phải phát triển các tiêu chuẩn mã hóa khách quan. Hệ mã hóa cử động khuôn mặt - FACS [37] là một trong những hệ thống mã hóa diễn đạt khuôn mặt được sử dụng rộng rãi nhất trong khoa học về hành vi.

FACS được phát triển bởi Ekman và Friesen nhằm mục đích xác định tất cả các cử động khuôn mặt có thể phân biệt được bằng mắt; nó tạo ra một mô tả rõ ràng, súc tích cho việc hoạt hóa các cơ của một diễn đạt khuôn mặt. 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 (một hay nhiều) cơ bên dưới tạo nên sự biến đổi đó được gọi là các *đơn vị cử động - Action Unit (AU)*; mỗi đơn vị cử động là một cử động cơ bản, được mô tả là kết hợp của một hoặc một số cơ trên khuôn mặt. 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 đơn vị cử động 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. Ví dụ, AU 4 mã hóa sự co của hai cơ có tên tiếng Anh là Corrugator supercilii và Depressor supercilii, đây là hai cơ mà khi co sẽ khiến lông mày hạ xuống. Đơn vị cử động này thường được thấy ở biểu cảm của các cảm xúc buồn, sợ, và giận. Như vậy, FACS cung cấp một ngôn ngữ khách quan và toàn diện cho việc mô tả các biểu cảm khuôn mặt và gắn kết chúng trở lại với những ý nghĩa đã được biết đến từ các nghiên cứu thuộc khoa học hành vi. FACS được ứng dụng rất phổ biến trong lĩnh vực tạo chuyển động cho khuôn mặt (ví dụ [3, 47]. Các biểu cảm thể hiện cảm xúc trên khuôn mặt được xây dựng dựa trên các đơn vị cử động của FACS. Phụ lục 1 của luận án mô tả đầy đủ danh sách các đơn vị cử động của

FACS.

2.2.2 Cảm xúc và giọng nói

Thực tế và các nghiên cứu cho thấy, sau cử động khuôn mặt, tiếng nói là kênh quan trọng thứ hai trong việc thể hiện các trạng thái cảm xúc. Theo [24], 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. Sự biến đổi giọng điệu thường được xem như thành phần phi lời nói, bởi vì những biến đổi này thực chất có mối quan hệ với cảm xúc giống như các kênh phi lời nói khác. Nội dung (bao gồm cả văn phạm) mà một người nói có thể hoặc không thể hiện đúng cảm xúc nhưng giọng điệu của người đó thường được xem như một thể hiện chính xác hơn về những gì mà anh ta/cô ta đang cảm thấy. Khi có sự mâu thuẫn giữa đặc trưng nội dung và đặc trưng giọng điệu thì người nghe thường chú ý hơn tới giọng điệu và không để ý đến thông điệp lời nói. Ví dụ, Mehrabian [99] đã chỉ ra rằng mỉa mai là nỗ lực có chủ ý để truyền tải một thông điệp thân thiện, vui vẻ với giọng điệu hằn học, vừa phủ nhận vừa làm thay đổi ý nghĩa của thông điệp. Một ông chồng có thể xem thường bà vợ với khuôn mặt đang đầy kem và tóc quấn lô bằng cách nói "dĩ nhiên là nhìn em rất đẹp" với giọng điệu châm biếm, mỉa mai. Một người cũng có thể làm trái lại kiểu trên, đó là nói điều gì đó khó chịu (ví dụ "đồ đáng ghét") với một giọng điệu vui vẻ và thông điệp sẽ được xem như là vui vẻ, hài lòng bởi vì thành phần giọng điệu lấn át nội dung lời nói.

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. Scherer [126] đã cố gắng chỉ ra ảnh hưởng của sự thay đổi trong độ căng cơ đối với việc phát âm. Tương tự, Ohala [109] đã đề cập đến ba sự biến đổi trong cơ thể có ảnh hưởng đến âm thanh được tạo ra trong quá trình phát âm. Ba sự biến đổi này là: Sự khô miệng hoặc thanh quản, tốc độ thở tăng, sự rung của cơ. Theo Abercrombie [2], nhịp điệu lời nói về bản chất là nhịp điệu cơ; không có ranh giới rõ ràng giữa hành vi lời nói và hành vi phi lời nói. Lời nói thực chất là hoạt động cơ và nó được quy về cùng loại ảnh hưởng như các hoạt động cơ khác. Hiệu ứng đáng chú ý nhất được tạo bởi hoạt động hay kích thích, hiệu ứng này bao gồm ba khía cạnh: *sự tích cực*, *độ căng*, và *sự cân bằng*. Những khía cạnh này biến đổi cùng nhau và

quyết định mẫu phát âm tổng thể. Ở mức đơn giản nhất, *sự tích cực* làm tăng độ to và tốc độ lời nói. Lời nói gắn với các cảm xúc tích cực thì to hơn và nhanh hơn (đó là tích cực hơn). Khía cạnh này biến lời nói nhẹ nhàng, chậm gắn với buồn và phiền muộn thành lời nói to, nhanh đi cùng với sự giận dữ và kích động cực độ. Mẫu thứ hai dựa trên sự tăng của độ căng cơ, sự tăng này khiến cao độ tăng, độ vang giảm, và giọng nói bị đứt gãy. Cao độ được dựa trên độ căng của dây thanh, độ căng này quyết định tốc độ rung. Độ căng tăng với các cảm xúc tích cực và điều này khiến tốc độ rung tăng lên, làm cho giọng nói có cao độ cao hơn. Độ vang được dựa trên độ mở của dây thanh. Với các cảm xúc tích cực, dây thanh trở nên căng và đóng, khiến cho giọng nói kém vang. Biến đổi ở độ căng cũng khiến cho giọng nói bị đứt gãy. Cuối cùng, hoạt động ảnh hưởng tới sự cân bằng của giọng nói. Cả lời nói và cử động cơ thể trở nên vụng về và không cân bằng khi kích thích ở mức cao. Giọng nói bị lạc điệu, độ vang trở nên không đều, và giọng bị đứt gãy. Điều quan trọng cần chú ý là mặc dù mỗi một trong số ba sự thay đổi trên là riêng biệt và rời rạc nhưng đều dựa trên cấu trúc giải phẫu của dây thanh và ảnh hưởng của sự kích thích lên dây thần kinh vận động. Mặc dù độ to, tốc độ, thanh điệu, và độ vang có sự biến đổi khác nhau giữa người này so với người khác xuất phát từ sự khác nhau về mặt sinh lý, nhưng trong các cảm xúc tích cực, những biến đổi này luôn diễn ra với cùng chiều hướng. Lời nói trở nên to hơn, nhanh hơn, kém vang hơn, và cao độ lớn hơn. Điều này dẫn tới gợi ý về cơ sở bẩm sinh cho các đặc trưng âm điệu mà con người không có hoặc có rất ít sự kiểm soát trên đó.

Tính bẩm sinh của biểu cảm giọng điệu có thể được hỗ trợ bởi nhiều tiêu chuẩn giống như các tiêu chuẩn được sử dụng để xác thực tính bẩm sinh của biểu cảm khuôn mặt. Các tiêu chuẩn này bao gồm sự hiện diện của các mẫu tương tự ở người trưởng thành, giữa các loài, và giữa các nền văn hóa (hay ngôn ngữ). Sự tăng cao độ và độ to trong các trạng thái cảm xúc tích cực được thấy ở nhiều loài. Ví dụ Jay [75] đã mô tả thay đổi ở cao độ giữa tiếng kêu của khỉ khi bình thường và khi đau đớn; và Andrews [7] đã đưa ra chú ý rằng nhiều động vật linh trưởng (và các động vật khác) có tiếng kêu biến đổi nhiều về mặt cường độ. Âm thanh với cường độ thấp gắn với tụ họp chào mừng và ăn uống, âm thanh với cường độ cao xuất hiện khi thất bại hoặc cảnh báo các đối tượng lạ. Malatesta [96] đã tổng hợp các nghiên cứu về biểu cảm âm điệu ở cả người và

động vật linh trưởng khi còn nhỏ, bà đưa ra kết luận rằng những biểu cảm này có thể là phổ biến và bẩm sinh. Sự nhạy cảm đối với các đặc trưng giọng điệu cũng phát triển rất sớm ở trẻ nhỏ. Trẻ sơ sinh phản ứng với thay đổi ở cao độ và độ to ngay khi mới được 6 tuần tuổi [13] và phản ứng lại với các mẫu phức tạp hơn khi được 6 tháng tuổi [14]. Những phản ứng sớm này chủ yếu dựa trên các đặc trưng diễn tả cảm xúc, và ví lý do này những gì được nói kém quan trọng hơn nhiều so với giọng điệu nói. Một số nghiên cứu về giao thoa văn hóa đã chỉ ra rằng con người có thể nhận diện cảm xúc trong lời nói với ngôn ngữ không quen thuộc [12, 123]. Các tác giả nhận thấy sự khác biệt lớn giữa các nhóm mà họ gọi là "thành thạo" và "không thành thạo" khi phán đoán các dấu hiệu khuôn mặt và cơ thể, nhưng rất ít sự khác biệt được thấy đối với các dấu hiệu giọng điệu. Thực tế, trong nhiều nghiên cứu, nhóm không thành thạo bao gồm trẻ em, người nói tiếng nước ngoài, bệnh nhân tâm thần, thực hiện rất tốt việc nhận diện cảm xúc thông qua các dấu hiệu giọng điệu. Các tác giả gợi ý rằng khả năng phán đoán cảm xúc thông qua các dấu hiệu giọng điệu phát triển sớm hơn khả năng phán đoán cảm xúc thông qua biểu cảm khuôn mặt và cử động cơ thể, và thậm chí có thể là bẩm sinh. Sự hiện diện của các mẫu giọng điệu tương tự ở trẻ em, giữa các loài, và giữa các ngôn ngữ cung cấp một số hỗ trợ cho niềm tin rằng các đặc trưng giọng điệu là bẩm sinh. Điều này dường như cũng sinh ra từ mối liên hệ gần gũi giữa biến đổi giọng điệu và biến đổi sinh lý; trong mối liên kết này, nhiều sự biến đổi trong giọng điệu của một người là do các mẫu bên dưới gắn với độ căng và sự cân bằng. Tuy nhiên, những quan sát này cần được kiểm định lại bởi vì các ngôn ngữ khác nhau sử dụng các đặc trưng giọng điệu khác nhau để truyền tải ý nghĩa và các đặc trưng biểu cảm có thể thay đổi giữa các ngôn ngữ.

Cuối cùng, theo [24], 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. Biến đổi giọng điệu cũng bị ảnh hưởng bởi các nhân tố tình huống như kích thước phòng, không gian, và âm thanh. Trong tất cả các trường hợp, biến đổi của đặc trưng giọng điệu xung quanh một chuẩn, như tăng độ to, cao độ, độ vang, và tốc độ sẽ có ý nghĩa hơn là giá trị tuyệt đối. Tất nhiên điều này không phải chỉ đúng duy cho nhất đặc trưng giọng điệu, bởi vì các chuẩn cũng tồn tại cho các kênh biểu cảm khác như tầm nhìn mắt và kiểm soát không gian cá nhân.

2.3 Cung cấp cảm xúc cho nhân vật ảo

Nhận ra tầm quan trọng của cảm xúc đối với chức năng nhận thức của con người, Picard [115] đã kết luận rằng nếu chúng ta muốn máy tính thực sự thông minh và tương tác với chúng ta một cách tự nhiên thì chúng cần phải có khả năng mô hình hóa, nhận dạng, và thể hiện cảm xúc. Trong lĩnh vực nghiên cứu về nhân vật ảo, cảm xúc nhận được nhiều sự quan tâm bởi ảnh hưởng của nó trong việc tạo các nhân vật ảo tin cậy(ví dụ [11, 19, 42]. Câu hỏi đặt ra là làm thế nào để cung cấp cảm xúc cho nhân vật ảo? Theo Thomas và Johnston [142] trạng thái cảm xúc của nhân vật ảo cần phải được định nghĩa một cách rõ ràng và được thể hiện tốt. Như vậy có hai vấn đề cần quan tâm khi giải quyết bài toán cung cấp cảm xúc cho nhân vật ảo: thứ nhất là *cung cấp trạng thái cảm xúc* cho nhân vật ảo, thứ hai là *cung cấp cơ chế thể hiện cảm xúc* cho nhân vật ảo. Nhằm cải tiến sự tương tác giữa người và máy tính, một nhân vật ảo trong máy tính có thể thể hiện biểu cảm khi mà cảm xúc không thực sự tồn tại bên trong nó. Tuy nhiên, việc này không cung cấp một cơ chế nhất quán cho việc thể hiện cảm xúc, khiến cho nhân vật ảo trở nên khó hiểu và kém thuyết phục. Ngược lại, khi nhân vật ảo đã được cung cấp trạng thái cảm xúc nhưng cơ chế thể hiện cảm xúc không tốt cũng sẽ khiến nhân vật ảo kém tự nhiên. Vì vậy, cách thức hiệu quả nhất đó là sử dụng các kỹ thuật mô hình hóa cho việc cung cấp trạng thái cảm xúc cũng như việc 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. Dựa trên các thuyết về cảm xúc, đặc biệt là thuyết đề xuất bởi Ortony cùng cộng sự [110] và thuyết đề xuất bởi Roseman [122], nhiều mô hình cảm xúc trên máy tính đã được phát triển. Các mô hình này được đề xuất ở nhiều dạng thức: hệ thống dựa trên luật [120, 42], hệ thống dựa trên luật mờ [41], hệ thống phân tán [147], hệ thống liên kết [92, 66], hệ thống dựa trên kế hoạch (plan based system) [54]... 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 đề xuất bởi Bui và cộng sự [19] đã giải quyết được các vấn đề vừa nêu.

Trong [19], các tác giả đã đề xuất ParleE - một hệ thống cung cấp trạng thái cảm xúc cho nhân vật ảo trên máy tính một cách linh động. ParleE thẩm định các sự kiện dựa trên việc học và một giải thuật lập lịch. ParleE cũng mô hình hóa cá tính và trạng thái động cơ, cũng như vai trò của chúng trong việc quyết định cách mà nhân vật ảo trải nghiệm cảm xúc. 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. Những nghiên cứu này có thể được chia thành hai lớp: phương pháp thể hiện cảm xúc tĩnh, và phương pháp thể hiện cảm xúc động. Phương pháp thể hiện cảm xúc tĩnh [4, 81, 83, 118] không có khả năng thể hiện các trạng thái cảm xúc liên tục; nó không cung cấp một cơ chế nhất quán nào cho việc tạo các biểu cảm thể hiện cảm xúc trên khuôn mặt. Phương pháp thể hiện cảm xúc động [18, 80, 119, 138, 147, 156, 95] lưu lại sự thay đổi của cường độ cảm xúc theo thời gian, cung cấp một cơ chế nhất quán cho việc tạo các biểu cảm thể hiện cảm xúc khuôn mặt. Trong phương pháp này, biểu cảm khuôn mặt được tạo ra từ các trạng thái cảm xúc liên tục theo cơ chế ánh xạ trực tiếp. Trong mỗi khoảng nhỏ thời gian, trạng thái cảm xúc được ánh xạ trực tiếp thành biểu cảm, sau đó biểu cảm này được thể hiện trên khuôn mặt. Kênh biểu cảm thứ hai được quan tâm sau kênh khuôn mặt đó là kênh tiếng nói.

2.4 Kết chương

Chương 2 của luận án đã tổng kết các nghiên cứu tâm lý học liên quan đến cảm xúc, các nghiên cứu về mối quan hệ giữa cảm xúc và các kênh biểu cảm. Những nghiên cứu này chính là cơ sở cho các nghiên cứu sử dụng máy tính để mô phỏng biểu cảm thể hiện cảm xúc của nhân vật ảo. Trong các kênh biểu cảm thì khuôn mặt và tiếng nói là hai kênh nhận được sự quan tâm nhiều nhất. Theo quan điểm của chúng tôi, trong việc mô phỏng mối quan hệ giữa cảm xúc và cử động khuôn mặt, kết quả nghiên cứu thuộc *quan điểm cảm xúc cơ bản* là hữu ích nhất. Vì vậy, luận án đi theo *quan điểm cảm xúc cơ bản* để đề xuất mô hình mô phỏng biểu cảm khuôn mặt thể hiện cảm xúc liên tục; nội dung nghiên cứu này được trình bày trong Chương 3. Từ những bằng chứng rõ ràng về mối quan

hệ giữa cảm xúc và đặc trưng giọng điệu, luận án cũng tập trung vào kênh tiếng nói khi giải quyết bài toán thể hiện cảm xúc cho nhân vật ảo nói tiếng Việt; nội dung nghiên cứu này được trình bày trong Chương 4.

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. Các nghiên cứu so sánh đóng góp liên quan của các kênh biểu cảm thông qua việc đưa ra những thông tin trái chiều từ các nguồn khác nhau nhìn chung đều chỉ ra rằng người quan sát sử dụng thông tin từ dấu hiệu khuôn mặt nhiều hơn bất kỳ nguồn nào khác. Mehrabian [98] đã chỉ ra rằng trong giao tiếp trực tiếp người - người, chỉ có 7% thông điệp cảm xúc được truyền tải qua từ ngữ, 38% thông điệp được truyền tải qua yếu tố giọng điệu, và có tới 55% thông điệp được truyền tải thông qua biểu cảm khuôn mặt. Sự quan trọng của biểu cảm khuôn mặt trong việc xét đoán cảm xúc xuất phát từ nhiều nhân tố [24]. Trước tiên, khuôn mặt là một trong những kênh giao tiếp phi lời nói dễ thấy nhất trong quá trình tương tác thông thường. Thứ hai, khuôn mặt cũng là một trong những nguồn giàu thông tin nhất về trạng thái cảm xúc. Chuyển động của cơ thể dường như có mối liên hệ gần nhất với quan điểm cá nhân như sự yêu mến, trạng thái, và mức hài lòng. Trái lại, biểu cảm khuôn mặt dường như là kênh phi lời nói duy nhất đủ khả năng để diễn tả các cảm xúc riêng biệt cũng như quan điểm chung chung. Lý do thứ ba của việc biểu cảm khuôn mặt thường được sử dụng khi xét đoán cảm xúc là niềm tin rằng có một mối liên hệ trực tiếp hơn giữa biểu cảm khuôn mặt và cảm xúc mà nó truyền tải so với các kênh biểu cảm khác. Khuôn mặt liên quan chủ yếu tới biểu cảm thể hiện cảm xúc và là một nguồn thông tin đáng tin cậy ngay cả khi các dấu hiệu tình huống không rõ ràng hoặc mâu thuẫn.

Khuôn mặt là phần biểu cảm nhất của cơ thể trong việc thể hiện cảm xúc,

có vai trò thiết yếu trong giao tiếp của con người [36]. 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. Như đã trình bày trong Chương 2, các kết quả nghiên cứu đã chỉ ra rằng tồn tại một mối liên kết giữa cử động trên khuôn mặt và trạng thái cảm xúc [28]. 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 con người. Cho tới nay, có nhiều nghiên cứu về mối quan hệ này đã được công bố (ví dụ [37, 38, 50, 111, 30, 143]). Tuy nhiên, hầu hết các nghiên cứu tập trung vào việc phân tích mối quan hệ nhưng lại không xem xét nó cùng với các yếu tố thời gian. Bên cạnh đó, các kỹ thuật tạo biểu cảm khuôn mặt thể hiện cảm xúc cũng đã được đề xuất (ví dụ [18, 58, 114, 137]). Những kỹ thuật này tập trung vào việc tạo biểu cảm khuôn mặt tĩnh từ cảm xúc; việc thể hiện trạng thái cảm xúc liên tục cho nhân vật ảo chưa được quan tâm nhiều, trừ nghiên cứu của tác giả Bui [16]. (Trạng thái cảm xúc liên tục có cường độ cảm xúc thay đổi liên tục theo thời gian; trái lại, trạng thái cảm xúc rời rạc là trạng thái mà trong một khoảng thời gian có một cảm xúc nào đó tồn tại với cường độ không đổi). Trong nghiên cứu [16], tác giả Bui và cộng sự đã đề xuất một cơ chế tạo biểu cảm khuôn mặt từ trạng thái cảm xúc liên tục. 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 ba chiều. Tuy nhiên, việc sử dụng ánh xạ trực tiếp như thế này sẽ tạo ra biểu cảm không tự nhiên khi có một trạng thái cảm xúc với cường độ cao xảy ra trong thời gian dài. Trong tình huống đó, biểu cảm có thể sẽ xuất hiện trên khuôn mặt trong thời gian khá dài; điều này có thể làm giảm tính tự nhiên của nhân vật ảo bởi vì theo kết quả nghiên cứu tâm lý và sinh lý học, một biểu cảm khuôn mặt thể hiện cảm xúc thường chỉ kéo dài trong khoảng từ 3 đến 4 giây [56].

Chương này của luận án đề xuất mô hình tạo biểu cảm khuôn mặt thể hiện trạng thái cảm xúc liên tục của nhân vật ảo. Mục tiêu của mô hình đề xuất là tạo ra biểu cảm tự nhiên, hạn chế được nhược điểm của các nghiên cứu đã công bố. Dựa trên kết quả nghiên cứu tâm lý và sinh lý học, luận án đề xuất mô hình thứ nhất cho bài toán thể hiện trạng thái cảm xúc liên tục của nhân vật ảo trên khuôn mặt. Ý tưởng chính của mô hình đó là một biểu cảm thể hiện cảm xúc trên khuôn mặt xuất hiện trong vài giây chỉ khi có sự thay đổi đáng

kể của trạng thái cảm xúc. Tiếp theo, dựa trên kết quả phân tích cơ sở dữ liệu video, luận án đề xuất mô hình thứ hai cho bài toán thể hiện trạng thái cảm xúc liên tục của nhân vật ảo trên khuôn mặt. Chúng tôi phân tích một cơ sở dữ liệu video để tìm câu trả lời cho câu hỏi: cử động khuôn mặt thể hiện cảm xúc thay đổi như thế nào theo thời gian. Mục tiêu là tìm các mẫu (theo thời gian) của cử động khuôn mặt cho sáu cảm xúc cơ bản; việc này nhằm mục đích cải tiến việc mô phỏng biểu cảm khuôn mặt thể hiện cảm xúc liên tục của nhân vật ảo. Để thực hiện mục tiêu này, trước tiên các kỹ thuật nhận dạng biểu cảm khuôn mặt được sử dụng để tự động phân tích một cơ sở dữ liệu video tự nhiên. Giả thuyết chúng tôi đưa ra là: khi một cảm xúc được kích hoạt, biểu cảm khuôn mặt tương ứng thể hiện cảm xúc xảy ra theo chuỗi với cường độ giảm dần. Ví dụ, khi có một sự kiện xảy ra, kích hoạt cảm xúc vui của một người, anh ta/cô ta sẽ không cười với cường độ lớn trong suốt khoảng thời gian mà cảm xúc vui tồn tại. Thay vào đó, anh ấy/cô ấy sẽ thể hiện một chuỗi các biểu cảm cười với cường độ giảm dần. Để chứng thực giả thuyết này, chuyển động của các điểm đặc trưng trên các khuôn mặt trong cơ sở dữ liệu được xác định và so khớp một cách tự động với các "mẫu" được định nghĩa trước. Dựa trên các mẫu theo thời gian này, luận án đề xuất mô hình thứ hai cho việc thể hiện trạng thái cảm xúc liên tục của nhân vật ảo trên khuôn mặt. Các thực nghiệm đã được tiến hành để đánh giá hiệu quả của các mô hình đề xuất.

Nội dung của chương được tổ chức như sau. Phần 3.2 trình bày tóm tắt về các nghiên cứu liên quan. Tiếp theo, Phần 3.3 trình bày hai mô hình đề xuất cho bài toán tạo biểu cảm khuôn mặt thể hiện trạng thái cảm xúc liên tục của nhân vật ảo. Sau đó, thực nghiệm và đánh giá được trình bày ở Phần 3.4.

3.2 Những nghiên cứu liên quan

Như đã trình bày trong Chương 2, theo quan điểm của 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. Hầu hết các nghiên cứu tâm lý học thuộc quan điểm này đều liên quan đến việc định nghĩa tập các loại biểu cảm. Có lẽ, nghiên cứu được biết đến nhiều nhất và hay được sử dụng nhất trong phân loại biểu cảm khuôn mặt chính là nghiên cứu giao thoa văn hóa

về sự tồn tại của các loại biểu cảm nhất quán [36, 34, 70]. Ekman đã định nghĩa sáu loại biểu cảm như vậy; sáu loại biểu cảm này được gọi là *sáu cảm xúc cơ bản*: vui, buồn, ngạc nhiên, sợ, giận, và khinh bỉ [36]. Ông mô tả mỗi cảm xúc cơ bản dưới dạng một biểu cảm khuôn mặt nhất quán và duy nhất, biểu cảm này đặc trưng cho cảm xúc đó. 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. Nghiên cứu sâu rộng trong [37] đã chỉ ra rằng sự kết hợp một số đơn vị cử động có mối liên kết với sáu mẫu khuôn mặt "phổ quát" của các cảm xúc cơ bản. EMFACS [50] được đề xuất bởi Friesen và Ekman tương tự như hệ mã hóa cử động khuôn mặt FACS nhưng nó chỉ quan tâm đến các cử động khuôn mặt liên quan đến cảm xúc. Ekman và Hager [38] cũng đã đưa ra một cơ sở dữ liệu được gọi là FACSAID (viết tắt của "facial action coding system affect interpretation database") cho phép chuyển cảm xúc liên quan tới FACS thành ý nghĩa tình cảm. Trong nghiên cứu [30], tất cả các ảnh trong một cơ sở dữ liệu bao gồm hình ảnh khuôn mặt ở trạng thái không cảm xúc, ở sáu cảm xúc cơ bản, và ở 15 cảm xúc phức phức hợp đã được mã hóa FACS nhằm phân tích mối quan hệ giữa cảm xúc và cử động khuôn mặt. Tian và cộng sự [143] đã phát triển một hệ thống phân tích khuôn mặt tự động; hệ thống này có thể nhận diện một cách tự động sáu AU ở phần mặt trên và mười AU ở phần mặt dưới, giúp cho việc xác định các biểu cảm và từ đó xác định trạng thái cảm xúc.

Hầu hết các nghiên cứu đã được công bố cố gắng xử lý các cảm xúc cơ bản và một số nghiên cứu xử lý các cảm xúc không phải là cơ bản. Tuy nhiên, theo hiểu biết của chúng tôi, hầu như chưa có nghiên cứu nào xem xét động thái theo thời gian của 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. Những thuật ngữ quan trọng liên quan đến động thái thời gian được sử dụng là: *onset*, *apex*, và *offset* [40]. *Onset* là thời điểm mà biểu cảm khuôn mặt bắt đầu xuất hiện,*apex* là thời điểm mà biểu cảm ở đỉnh của nó, và *offset* là thời điểm biểu cảm khuôn mặt biến mất. Tương tự như vậy, *khoảng thời gian onset* (onset-duration) được định nghĩa là thời gian từ khi bắt đầu biểu cảm cho tới đỉnh, *khoảng thời gian apex* (apex - duration) là tổng thời gian ở đỉnh, và *khoảng thời gian offset* (offset - duration)là tổng thời gian từ đỉnh cho tới khi dừng biểu cảm. Pantic và Patras đã thành công trong việc nhận diện các AU trên khuôn mặt và phân khúc thời gian của chúng [111]. Từ đó, có thể nhận diện một số lượng lớn các biểu cảm. Tuy

nhiên, nghiên cứu này chỉ phân tích các AU đơn lẻ, nó không đề cập đến "mẫu" biểu cảm khuôn mặt cho các cảm xúc theo miền thời gian. Trong Phần 3.3.2.1 chúng tôi thực hiện phân tích trên cơ sở dữ liệu video khuôn mặt để tìm ra các "mẫu" này.

Từ hiểu biết về mối quan hệ giữa cảm xúc và cử động khuôn mặt, trên cơ sở có khả năng mô phỏng mối quan hệ này, đến nay, 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: các phương pháp thể hiện cảm xúc tĩnh, và các phương pháp thể hiện cảm xúc động.

Phương pháp thể hiện cảm xúc tĩnh: Nhiều nhà nghiên cứu gồm Albrecht [4], Kulander [81], Latta [83], Raouzaiou [118]đã sử dụng mô hình bánh xe cảm xúc (wheel) được mô tả bởi Plutchik [117] để phát triển các hệ thống hoạt hóa khuôn mặt. Mô hình bánh xe cảm xúc này cho phép các nhà nghiên cứu tạo ra các cơ chế để ánh xạ các trạng thái cảm xúc thành các biểu cảm khuôn mặt được nhận ra 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 thể hiện 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 trước đó của khuôn mặt. Đây thực sự là một điểm yếu đáng kể. 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 (ví dụ vui) thành một cảm xúc trái ngược (ví dụ giận dữ) chiếm một thời gian đáng kể, điều này không phù hợp lắm.

Phương pháp thể hiện cảm xúc động: Thể hiện cảm xúc động gồm các hệ thống của Bui và cộng sự [18], Kshirsagar và Magnenat Thalman [80], Mahardika và cộng sự [95], Reilly [119], Tanguy [138], Velásquez [147], Zhang và cộng sự [156]. 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 các biểu cảm thể hiện cảm xúc trên khuôn mặt và giải quyết được các giới hạn của các phương pháp thể hiện cảm xúc tĩnh.

Hệ thống do Kshirsagar và Magnenat Thalman [80] đề xuất dùng Bayesian Belief Network thể hiện trạng thái cảm xúc để lựa chọn biểu cảm tiếp theo sẽ được hiển thị trên khuôn mặt của nhân vật ảo. Đầu vào của hệ thống là một

danh sách các phản ứng có thể, gắn với khả năng trạng thái cảm xúc được cung cấp bởi một hệ thống robot gọi là ALICE. Với mỗi khả năng trạng thái cảm xúc, hệ thống sẽ tính toán xác suất của mỗi trạng thái cảm xúc có thể, trong mối quan hệ với cá tính của nhân vật ảo và trạng thái cảm xúc trước đó. Trạng thái cảm xúc với xác suất cao hơn sẽ được chọn. Nhân vật ảo có thể ở một trong 24 trạng thái cảm xúc, các trạng thái này được ánh xạ tới một trong sáu biểu cảm khuôn mặt được nhận diện là nhất quán. Một số trạng thái cảm xúc có thể có cùng biểu cảm khuôn mặt. Ưu điểm của phương pháp này là dùng các trạng thái trước đó để lựa chọn các biểu cảm tiếp theo. Tuy nhiên, số các biểu cảm có thể được hiển thị bởi hệ thống này bị giới hạn.

Dựa trên mô tả của Picard về cường độ cảm xúc và các bộ lọc cảm xúc [115], Tanguy [138] đã phát triển một mô hình thể hiện cảm xúc động. Mô hình này có thể thể hiện số kiểu trạng thái bất kỳ, như các cảm xúc, drives,... Kiến trúc của mô hình gồm hai lớp, tác giả đặt tên là Secondary Emotions và Moods, mỗi lớp thể hiện một loại cảm xúc khác nhau. Trạng thái cảm xúc thuộc lớp Secondary Emotions được mô tả bởi sáu biến tương ứng với sáu cảm xúc cơ bản: vui, buồn, giận dữ, ngạc nhiên, khinh bỉ, sợ hãi. Trạng thái cảm xúc thuộc lớp Mood được mô tả bởi hai biến: Energy và Tension. Giá trị của hai biến này thể hiện cá tính cảm xúc của nhân vật ảo. Để diễn tả trạng thái Secondary Emotions, cảm xúc với cường độ cao nhất được chọn và biểu cảm khuôn mặt tương ứng với cảm xúc đó được tạo. Với trạng thái thuộc lớp Moods, biểu cảm khuôn mặt được tạo tương ứng với cường độ được biểu diễn bởi giá trị của biến Tension. Sau đó, biểu cảm khuôn mặt cuối cùng sẽ thể hiện trên nhân vật ảo được tạo ra bằng cách chọn mức co cơ cao nhất từ các biểu cảm được tạo từ trạng thái của hai lớp Secondary Emtions và Moods.

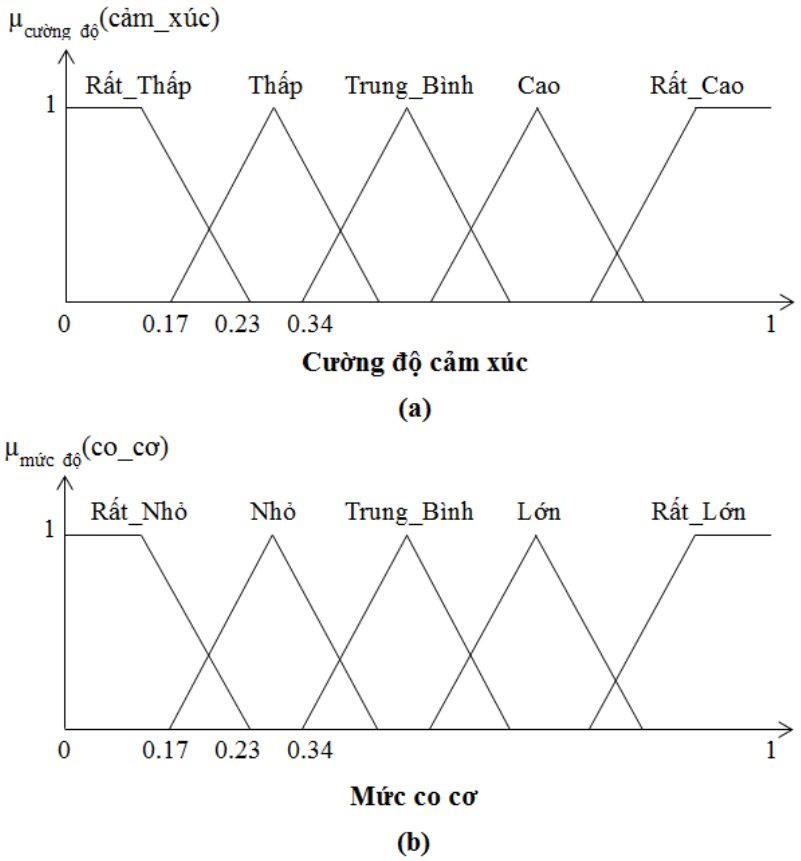
Đi theo quan điểm đa chiều về mối quan hệ giữa cảm xúc và biểu cảm khuôn mặt, Zhang và cộng sự đã đề xuất một framework đa lớp để tổng hợp biểu cảm khuôn mặt thể hiện cảm xúc cho nhân vật ảo MPEG4 [156]. Đầu vào của hệ thống là bộ ba tham số "pleasure-arousal-dominance" (PAD) thể hiện trạng thái cảm xúc của nhân vật ảo, đầu ra cử động khuôn mặt thể hiện trạng thái cảm xúc đầu vào. Để tạo cử động khuôn mặt thể hiện cảm xúc, hệ thống không thao tác trực tiếp trên các tham số hoạt hóa khuôn mặt (FAPs) của MPEG4; thay vào đó, một tập các tham số biểu cảm thành phần (PEPs) được

đưa ra để mô tả biểu cảm khuôn mặt cục bộ (như mở mắt, nhướn mày,...). Tập PEPs mô hình hóa sự tương quan giữa các FAPs bên trong một vùng khuôn mặt cục bộ. Để tổng hợp biểu cảm khuôn mặt, trước tiên bộ ba PAD được ánh xạ thành các tham số PEPs; sau đó, các tham số PEPs được chuyển thành tham số FAPs; cuối cùng, các tham số FAPs được dùng để hoạt hóa khuôn mặt, tạo biểu cảm khuôn mặt tương ứng với cảm xúc đầu vào.

Nghiên cứu được đề xuất bởi Bui và cộng sự [18] dùng hệ thống dựa trên luật mờ để ánh xạ trạng thái cảm xúc thành biểu cảm khuôn mặt: một hệ thống con có tên "Single Expression Mode FRBS" được sử dụng để tạo biểu cảm khuôn mặt từ một cảm xúc đơn, hệ thống con còn lại có tên "Blen Expression Mode FRBS" được sử dụng để tạo biểu cảm khuôn mặt từ hai cảm xúc với cường độ cao nhất. Đầu vào của hệ thống dựa trên luật mờ là cường độ sáu cảm xúc cơ bản (giá trị số thực), đầu ra của hệ thống là mức co cơ mặt (giá trị số thực) để tạo cử động diễn tả cảm xúc đầu vào. Sáu cảm xúc cơ bản được được xử lý bao gồm vui, buồn, ngạc nhiên, sợ hãi, khinh bỉ, và giận dữ. Nếu độ chênh lệch giữa hai cường độ cảm xúc đầu vào cao nhất lớn hơn 0.5 thì chỉ có một cảm xúc được thể hiện, và hệ thống con "Single Expression Mode FRBS" được kích hoạt. Trái lại, nếu độ chênh lệch này nhỏ hơn 0.5 thì có hai cảm xúc được thể hiện đồng thời trên khuôn mặt, hệ thống con "Blen Expression Mode FRBS" sẽ được kích hoạt. Khi có hai cảm xúc được thể hiện, chúng được diễn tả trên các vùng khác nhau của khuôn mặt; cơ chế này giải quyết được vấn đề xung đột có thể xảy ra vì mỗi cảm xúc được thể hiện ở một vùng riêng biệt.

Bui và cộng sự sử dụng 5 tập mờ *Rất_ Thấp, Thấp, Trung_ Binh, Cao, Rất_ Cao* để mô hình hóa cường độ cảm xúc, và 5 tập mờ *Rất_ Nhỏ, Nhỏ, Trung_ Binh, Lớn, Rất_ Lớn* để mô tả mức co cơ mặt. Các *hàm thành viên* (membership function) và *hỗ trợ* (support) được xác định bằng tay qua thực nghiệm, được minh họa trên Hình 3.1. Sau đó, dựa trên nghiên cứu về cử động khuôn mặt thể hiện cảm xúc [36, 37], tác giả định nghĩa hai tập luật mờ *if-then* ghi lại mối quan hệ giữa cường độ cảm xúc và mức co cơ mặt; một tập cho hệ thống con Single Expression Mode FRBS và một tập cho hệ thống con Blen Expression Mode FRBS. Các luật mờ *if-then* có dạng như sau:

> **If** cường độ cảm xúc vui là *Rất_ Thấp* **then** mức co của cơ Zygomatic Major là *Rất_ Nhỏ*, mức co của cơ Zygomatic Minor là *Rất_ Nhỏ*,...



Hình 3.1: (a): Hàm thành viên cho cường độ cảm xúc. (b): Hàm thành viên cho mức co cơ [18].

Hệ thống dựa trên luật mờ chuyển cường độ cảm xúc thành mức co cơ mặt theo các bước như sau: trước tiên, cường độ cảm xúc đầu vào được mờ hóa; sau đó, các luật mờ *if-then* thích hợp được áp dụng để chuyển cường độ cảm xúc (giá trị mờ) thành mức co cơ (giá trị mờ); cuối cùng, mức co cơ mặt được khử mờ bằng phương pháp *Trung tâm vùng* (Center of Area). Ưu điểm của phương pháp đề xuất bởi Bui và cộng sự là việc sử dụng hệ thống dựa trên luật mờ cho phép kết hợp các mô tả định tính như "rất rất vui" với các thông tin định lượng như cường độ cảm xúc hay mức co cơ.

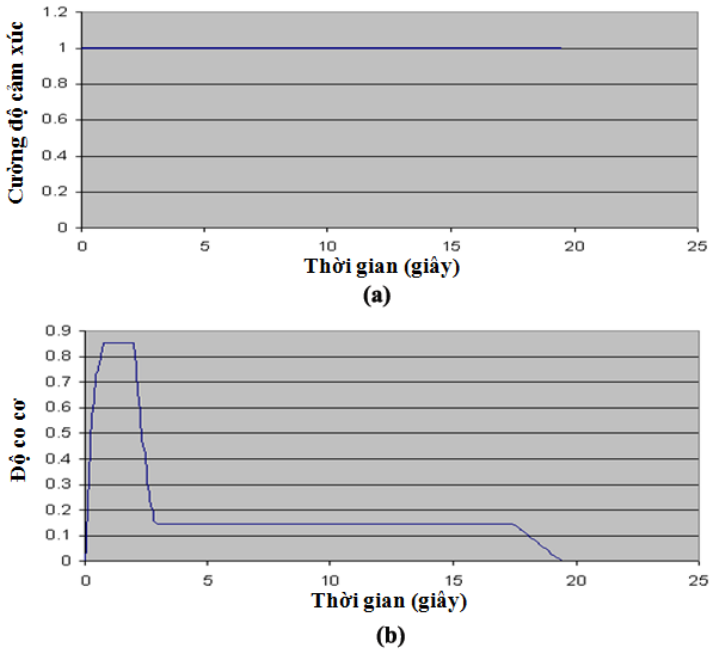
Có thể thấy rằng các phương pháp thể hiện cảm xúc động hiệu quả hơn các phương pháp thể hiện cảm xúc tĩnh. Tuy nhiên, các hệ thống thể hiện cảm xúc độrng hiện có mới chỉ giải quyết việc thể hiện cảm xúc mà chưa tính đến yếu tố thời gian của các biểu cảm. Các phương pháp này đều mới chỉ thực hiện ánh xạ

trực tiếp từ cảm xúc đã biết thành biểu cảm khuôn mặt tương ứng. Trong thực tế, cảm xúc của con người rất phức tạp, và thường thì chúng ta không thể biết trước cảm xúc của một người sẽ xuất hiện như thế nào. Trong trường hợp có một cảm xúc diễn ra trong một khoảng thời gian dài thì việc ánh xạ trực tiếp từ các cảm xúc thành các biểu cảm khuôn mặt sẽ làm giảm tính tự nhiên của các nhật vật ảo. Thực tế điều này rất hay xảy ra vì các cảm xúc có xu hướng triệt tiêu chậm hơn so với biểu cảm khuôn mặt. Theo kết quả nghiên cứu tâm lý và sinh lý học, thường thì một biểu cảm thể hiện cảm xúc trên khuôn mặt chỉ xuất hiện trong vài giây [56]. Mô hình luận án đề xuất trong Phần 3.3 sẽ giải quyết được hạn chế này.

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 [56], luận án đề xuất mô hình thứ nhất tạo biểu cảm thể hiện trạng thái cảm xúc liên tục của nhân vật ảo trên khuôn mặt. 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. Điều đó có nghĩa là các biểu cảm thể hiện cảm xúc trên khuôn mặt chỉ xuất hiện khi có một tác nhân kích thích đáng kể làm thay đổi trạng thái cảm xúc. Biểu cảm thể hiện cảm xúc sẽ không được giữ trên khuôn mặt trong một khoảng thời gian dài, trong khi đó cảm xúc lại triệt tiêu chậm. Tuy nhiên, các biểu cảm của tâm trạng có thể được giữ trên khuôn mặt trong khoảng thời gian dài hơn rất nhiều.

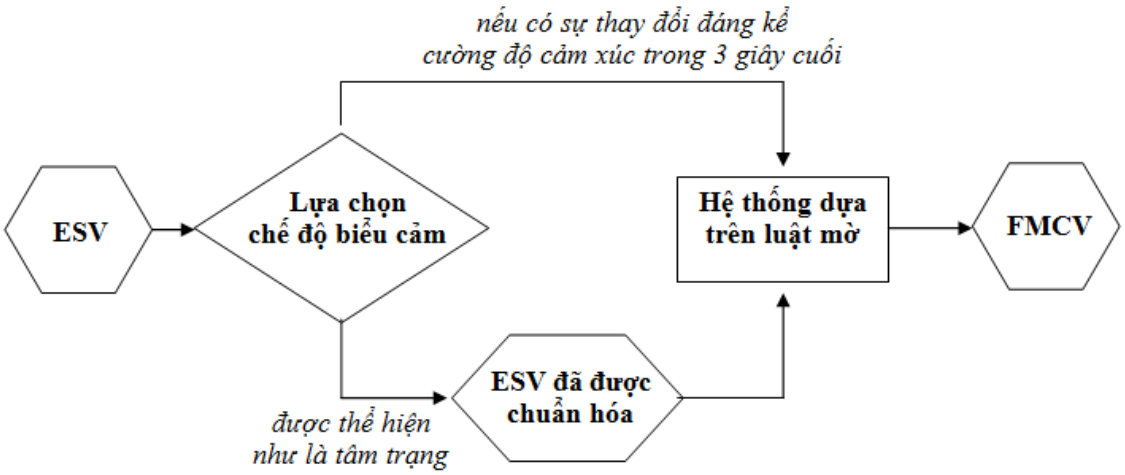


Hình 3.2: Ví dụ minh họa cơ chế của mô hình đề xuất thứ nhất chuyển cường độ cảm xúc thành mức co cơ.

Hình 3.2 chỉ ra một ví dụ của cơ chế tạo biểu cảm vừa được nêu. Hình vẽ phía trên (hình a) biểu diễn cảm xúc vui với cường độ cực đại là 1 kéo dài trong 20 giây, từ giây thứ 0. Nếu theo ánh xạ trực tiếp thì cảm xúc vui trong 20 giây này sẽ tạo ra độ co cơ Zymgomatic Major (cơ cười) tương ứng trong 20 giây và khiến cho khuôn mặt của nhân vật ảo "cười" trong 20 giây, hành động này không tự nhiên. Theo cơ chế của mô hình đề xuất thứ nhất, độ co cơ ứng với cảm xúc vui với cường độ 1 chỉ được giữ trong khoảng 3 giây đầu tiên, trong những giây còn lại độ co cơ sẽ được tạo ra ứng với "tâm trạng vui", có giá trị nhỏ hơn và không khiến cho khuôn mặt của nhân vật ảo "cười" trong thời gian dài.

Mô hình đề xuất thứ nhất dùng hệ thống trong [18] để chuyển một trạng thái cảm xúc tĩnh thành biểu cảm khuôn mặt. Như chỉ ra trên Hình 3.3, mô hình gồm bốn thành phần:

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 như [16]. Mỗi ESV là một véc tơ chứa cường độ của sáu cảm xúc tại thời điểm t, được biểu diễn bởi



Hình 3.3: Mô hình thứ nhất chuyển trạng thái cảm xúc liên tục thành biểu cảm khuôn mặt.

các số thực:

$$ESV^t = (e_1^t, e_2^t, \ldots, e_6^t), \tag{3.1}$$

trong đó $0 \leq e_i^t \leq 1$, $i = 1, ..., 6$.

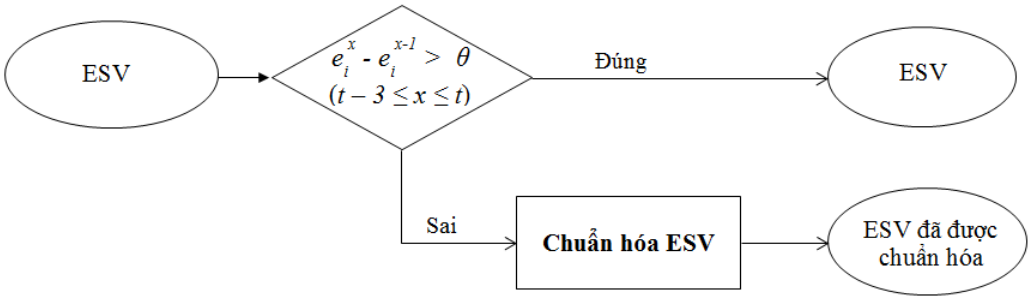
2. Đầu ra là chuỗi véc tơ độ co cơ mặt (FMCV) theo thời gian. Mỗi véc tơ FMCV tại thời điểm t được mô tả như sau:

$$FMCV^t = (m_1^t, m_2^t, \ldots, m_{19}^t), \tag{3.2}$$

trong đó $0 \leq m_i^t \leq 1$, $i = 1, ..., 19$. Đây là một véc tơ biểu diễn độ co của 19 cơ bên phía phải của mô hình khuôn mặt 3D trong [15].

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. Hoạt động của mô đun này được minh họa bởi sơ đồ khối trong Hình 3.4. Cụ thể, thành phần này sẽ thực hiện việc kiểm tra xem có sự tăng đáng kể trong cường độ của cảm xúc bất kỳ trong ba giây cuối (khoảng thời gian trung bình của một biểu cảm thể hiện cảm xúc), tức là nếu:

$$e_i^x - e_i^{x-1} > \theta, \tag{3.3}$$



Hình 3.4: Hoạt động của mô đun *Lựa chọn chế độ biểu cảm* trong mô hình đề xuất thứ nhất.

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 (ngưỡng θ được chọn giá trị 0.3 qua thực nghiệm). 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 [18]. Ngược lại, khi không có sự thay đổi đáng kể của cường độ cảm xúc, véc tơ EVS được chuẩn hóa về cường độ thấp hơn và sau đó được chuyển thành véc tơ FMCV cũng dùng hệ thống dựa trên luật mờ như trên. Theo cách này, cảm xúc được thể hiện như là tâm trạng - trạng thái ở cường độ thấp và thời gian kéo dài của các cảm xúc.

4. *Hệ thống dựa trên luật mờ* được dùng để chuyển véc tơ trạng thái cảm xúc (ESV) thành véc tơ độ co cơ mặt (FMCV); hệ thống này được đề xuất trong [18]. Như đã trình bày trong Phần 3.2, tác giả dùng 5 tập mờ để mô hình hóa cường độ cảm xúc, 5 tập mờ để mô tả mức co cơ, và định nghĩa hai tập luật mờ *if - then* để ghi lại mối quan hệ giữa cường độ cảm xúc và mức co cơ mặt. Quá trình chuyển cường độ cảm xúc thành mức co cơ được thực hiện qua ba bước: mờ hóa giá trị cường độ cảm xúc, áp dụng các luật mờ *if-then*, khử mờ giá trị mức co cơ.

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. Trước tiên, chúng tôi tiến hành phân tích một cơ sở dữ liệu video để tìm câu trả lời cho câu hỏi: cử động khuôn mặt

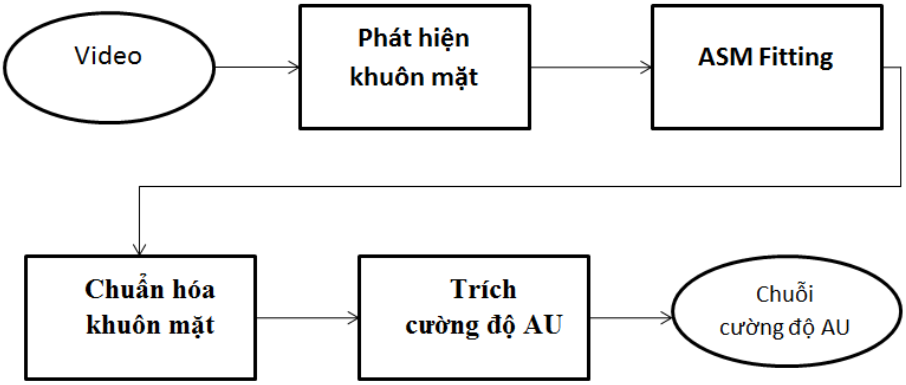
thể hiện cảm xúc thay đổi như thế nào theo thời gian. Sau đó, dựa trên các mẫu (theo thời gian) của cử động khuôn mặt thể hiện sáu cảm xúc cơ bản, mô hình thứ hai cho bài toán tạo biểu cảm khuôn mặt thể hiện trạng thái cảm xúc liên tục được xây dựng.

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; cơ sở dữ liệu này chứa các video được chọn từ ba cơ sở dữ liệu là MMI [112], FEEDTUM [152] và DISFA [102]. Cơ sở dữ liệu biểu cảm khuôn mặt MMI [112] chứa cả biểu cảm tự nhiên và biểu cảm "diễn", tuy nhiên luận án chỉ sử dụng các biểu cảm tự nhiên. Những biểu cảm này thuộc phần IV và phần V của cơ sở dữ liệu - hai phần chứa các video thể hiện sáu cảm xúc cơ bản. Đây là các video được thu thập từ quá trình thực nghiệm, trong đó các nhà nghiên cứu cho người tham gia thực nghiệm xem hình ảnh, video, hoặc clip ngắn của phim hoạt hình, chương trình hài kịch, hay âm thanh của các tác nhân tạo ra cảm xúc. Cơ sở dữ liệu FEEDTUM [152] chứa các cảm xúc được gợi ra một cách tự nhiên của 18 đối tượng. Với mỗi đối tượng, bênh cạnh trạng thái không cảm xúc, cơ sở dữ liệu này còn có các trạng thái cảm xúc giận dữ, khinh bỉ, sợ hãi, vui vẻ, ngạc nhiên, và buồn chán. Để có thể gợi ra cảm xúc một cách tự nhiên nhất, các video kích thích cảm xúc được lựa chọn cẩn thận và trình chiếu; phản ứng của người xem được ghi lại. Trong cơ sở dữ liệu DISFA [102], hình ảnh của 27 thanh niên được ghi lại trong khi họ xem các video clip có mục đích gợi ra biểu cảm thể hiện cảm xúc tự nhiên.

Từ ba cơ sở dữ liệu trên, chúng tôi chọn các video trong đó khuôn mặt người tham gia bắt đầu từ trạng thái không cảm xúc, tiến dần tới trạng thái đỉnh điểm của biểu cảm thể hiện cảm xúc, và sau đó trở lại trạng thái không cảm xúc. Vì mục tiêu là tìm "mẫu" theo thời gian của cử động khuôn mặt thể hiện cảm xúc nên những video được lựa chọn theo cách trên sẽ phù hợp cho quá trình phân tích. Cuối cùng có 215 video được chọn: 67 video cho cảm xúc vui, 25 video cho cảm xúc buồn, 25 video cho cảm xúc giận, 33 video cho cảm xúc khinh bỉ, 30 video cho cảm xúc sợ hãi, và 35 video cho cảm xúc ngạc nhiên. Các video này được xếp vào sáu danh mục, tùy theo cảm xúc mà khuôn mặt trong



Hình 3.5: Sơ đồ khối của hệ thống phân tích cử động khuôn mặt thể hiện cảm xúc.

video đó thể hiện.

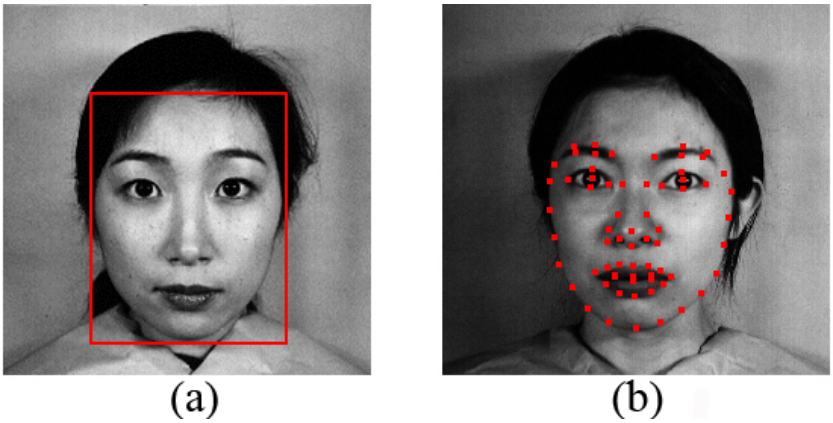
Phân tích cử động khuôn mặt thể hiện cảm xúc

Quá trình phân tích cử động khuôn mặt thể hiện cảm xúc được minh họa trong Hình 3.5. Đầu vào của hệ thống là một video, video này sẽ được xử lý theo từng frame. Với mỗi frame, mô đun *Phát hiện khuôn mặt* phát hiện khuôn mặt và trả về vị trí của nó. Sau đó, mô đun *ASM Fitting* thực hiện thao tác fitting và trả về ASM shape của khuôn mặt. Từ ASM shape này, mô đun *Chuẩn hóa khuôn mặt* tiến hành thao tác chuẩn hóa để đưa ASM shape một về kích thước chuẩn. Cuối cùng, mô đun *Trích cường độ AU* dùng các điểm đặc trưng thu được sau bước chuẩn hóa để trích ra cường độ của các AU liên quan tới sáu cảm xúc cơ bản. Cấu trúc và hoạt động chi tiết của các mô đun được trình bày trong phần tiếp theo.

A. Mô đun "Phát hiện khuôn mặt" (Face Detector):

Quá trình phân tích được bắt đầu với việc phát hiện khuôn mặt trong khung hình. Với mỗi frame của video đầu vào, mô đun *Phát hiện khuôn mặt* kiểm tra xem có tồn tại khuôn mặt người trong frame đó không, nếu có nó sẽ trả về kích thước và vị trí xấp xỉ của khuôn mặt được phát hiện. Trong nghiên cứu này, luận án sử dụng thuật toán Viola Jones [148] để phát hiện khuôn mặt. Kết quả của thuật toán phát hiện khuôn mặt được minh họa trong Hình 3.6(a).

B. Mô đun "ASM Fitting":



Hình 3.6: (a):Phát hiện khuôn mặt. (b): Các điểm đặc trưng trên khuôn mặt

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 mặ*, mô đun "ASM Fitting" sử dụng Active Shape Model [25] để tìm kiếm vị trí chính xác của các điểm đặc trưng trên khuôn mặt. Trong số các mô hình biến đổi khuôn mặt (deformable face models), chúng tôi sử dụng Active Shape Model (ASM) vì một số lý do. Thứ nhất, ASM có thể được xem là phương thức đơn giản nhất và nhanh nhất trong số các mô hình biến đổi, điều này phù hợp với nhu cầu cần phân tích một số lượng lớn frame từ nhiều video. Hơn nữa, vì tính đơn giản của nó, ASM được biết đến là mô hình có khả năng khái quát hóa tốt cho các đối tượng mới. Chúng tôi dùng vị trí của 68 điểm đặc trưng được đánh dấu bằng tay trên tập các ảnh tĩnh để huấn luyện mô hình ASM. Đầu ra của mô đun ASM Fitting là vị trí của 68 điểm đặc trưng trên khuôn mặt (được gọi là ASM shape), như được minh họa trong Hình 3.6(b).

C. Mô đun "Chuẩn hóa khuôn mặt" (Face Normalization):

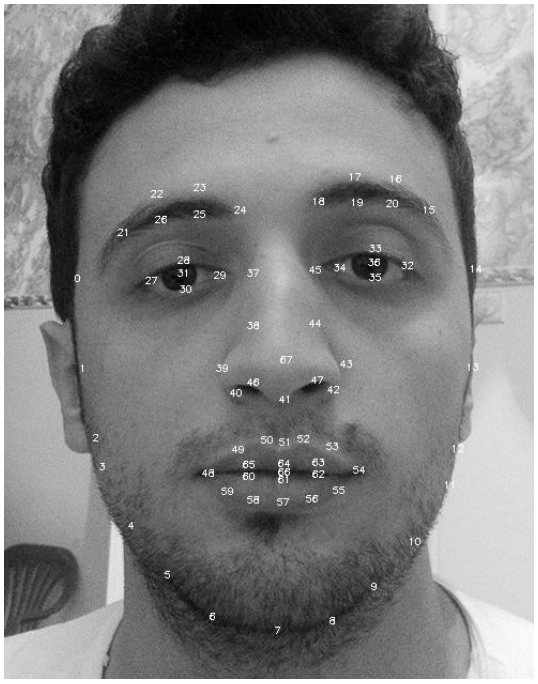
Do sự thay đổi vị trí máy quay, hay sự di chuyển của người tham gia thực nghiệm, kích thước khuôn mặt trong các frame của cùng một video có thể không giống nhau. Vì thế, ASM shape của các frame này cũng không có cùng kích thước. Điều này có thể dẫn tới sự kém chính xác trong kết quả phân tích. Vì vậy, cần phải có thao tác chuẩn hóa để đưa tất cả các ASM shape của một video về cùng một kích thước. Mô đun "Chuẩn hóa khuôn mặt" dùng khoảng cách giữa hai con ngươi mắt để thực hiện việc chuẩn hóa. Các ASM shape sẽ được chuẩn hóa sao cho sau khi thực hiện thao tác chuẩn hóa, khoảng cách giữa hai con ngươi mắt trong các ASM shape là bằng nhau.

D. Mô đun "Trích cường độ AU" (AUs Intensity Extractor):

Cảm xúc	**Action Unit**	**Đặc trưng khuôn mặt**
Vui	AU12	Kéo khóe môi (Lip Corner Puller)
	AU25	Tách môi trên và dưới (Lips Part)
Buồn	AU1	Nhướn mày trong (Inner Brow Raiser)
	AU4	Hạ lông mày (Brow Lowerer)
	AU15	Nén khóe môi (Lip Corner Depressor)
Sợ hãi	AU1	Nhướn mày trong (Inner Brow Raiser)
	AU2	Nhướn mày ngoài (Outer Brow Raiser)
	AU4	Hạ lông mày (Brow Lowerer)
	AU5	Nhướn mi trên (Upper Lid Raiser)
	AU20	Kéo căng môi (Lip Stretcher)
	AU25	Tách môi trên và dưới (Lips Part)
	AU26	Hạ hàm (Jaw Drop)
Khinh bỉ	AU15	Nén khóe môi (Lip Corner Depressor)
Giận dữ	AU4	Hạ lông mày (Brow Lowerer)
	AU5	Nhướn mi trên (Upper Lid Raiser)
	AU7	Căng mí mắt (Lid Tightener)
Ngạc nhiên	AU1	Nhướn mày trong (Inner Brow Raiser)
	AU2	Nhướn mày ngoài (Outer Brow Raiser)
	AU5	Nhướn mi trên (Upper Lid Raiser)
	AU25	Tách môi trên và dưới (Lips Part)
	AU26	Hạ hàm (Jaw Drop)

Bảng 3.1: Mô tả sáu cảm xúc cơ bản ([37, 30]).

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 Action Unit liên quan đến trạng thái cảm xúc được thể hiện trong video đầu vào. Luận án đi theo *quan điểm cảm xúc cơ bản* và dựa trên nghiên cứu của Ekman [37], Shichuan [30] về mối liên kết giữa tổ hợp AU và sáu mẫu khuôn mặt phổ quát của các cảm xúc vui, buồn, giận, khinh bỉ, sợ hãi, ngạc nhiên. Với mỗi cảm xúc, có một tập các AU liên quan, phân biệt nó với các cảm xúc khác, như được chỉ ra trong Bảng 3.1. Cường độ của các AU được tính thông qua việc xác định giá trị của đặc trưng khuôn mặt điển hình cho AU đó. Luận án định nghĩa 10 đặc trưng khuôn mặt tương ứng với một số điểm trong 68 điểm đặc trưng trên khuôn mặt; các điểm đặc trưng này được đánh số thứ tự như trong Hình 3.7.



Hình 3.7: Đánh số thứ tự các điểm đặc trưng trên khuôn mặt.

Các đặc trưng khuôn mặt điển hình cho các AU được mô tả trong Bảng 3.2; trong đó $d_{i,j}$ là khoảng cách giữa hai điểm đặc trưng i, j, $d_{i\_j,k}$ là khoảng cách giữa điểm i và đường thẳng nối hai điểm j, k, $l_{i,j,k,...}$ là chiều dài của đường xác định bởi các điểm $j, j, k, ...$

Kết quả phân tích

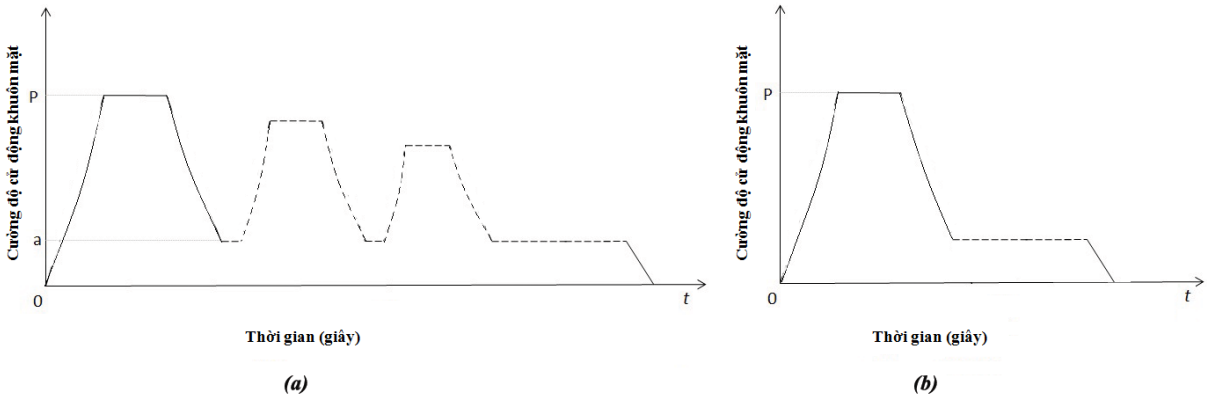
Với một video của mỗi loại cảm xúc, từ vị trí của các điểm đặc trưng trên khuôn mặt, cường độ của mỗi AU liên quan tới cảm xúc đó được tính lần lượt theo từng frame. Kết quả là chúng ta sẽ có một chuỗi các giá trị cường độ theo thời gian cho mỗi AU. Những chuỗi này được trích ra và dùng để vẽ đồ thị cường độ AU theo thời gian. Cuối cùng, các chuỗi và đồ thị của tất cả các AU từ tất cả các video thuộc một loại cảm xúc được dùng để tổng quát hóa mẫu theo thời gian cho biểu cảm khuôn mặt thể hiện cảm xúc đó. Từ việc quan sát các đồ thị, chúng tôi đưa ra giả thuyết rằng khi một cảm xúc được kích hoạt, biểu cảm khuôn mặt thể hiện cảm xúc đó xảy ra theo chuỗi với cường độ giảm dần. Từ đó, 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 3.8(a); mẫu theo thời gian cho biểu cảm thể hiện các cảm xúc khinh bỉ, giận, sợ, và ngạc nhiên được

Đặc trưng	**Mô tả**	**AU**
f1	$d_{18\_31,36} + d_{24\_31,36}$	AU1
f2	$d_{16\_31,36} + d_{17\_31,36} + d_{22\_31,36} + d_{23\_31,36}$	AU2
f3	$d_{15...20\_31,36} + d_{21...26\_31,36}$	AU4
f4	$d_{30,28} + d_{35,33}$	AU5
f5	$d_{30,28} + d_{35,33}$	AU7
f6	$l_{48,60,61,62,54}$	AU12
f7	$dy_{48,66} + dy_{54,66}$	AU15
f8	$d_{48,54}$	AU20
f9	$dy_{61,64}$	AU25
f10	$dy_{5...9,66}$	AU26

Bảng 3.2: Mô tả các đặc trưng khuôn mặt điển hình cho các AU.

chỉ ra trong Hình 3.8(b).

Trong các mẫu của biểu cảm, chúng ta thấy có phần liền nét và phần đứt nét. Sự khác nhau giữa hai phần này là phần liền nét thì luôn luôn xuất hiện còn phần đứt nét có thể xuất hiện, có thể không. Điều này có thể được giải thích như sau. Trạng thái cảm xúc là nguyên nhân của sự xuất hiện các cử động khuôn mặt, những cử động này sẽ tạo nên biểu cảm khuôn mặt thể hiện cảm xúc đó. Khi trạng thái cảm xúc xuất hiện trong một khoảng thời gian với cường độ đủ lớn, nó sẽ dẫn tới sự xuất hiện của cử động khuôn mặt, kéo theo là biểu cảm khuôn mặt trong khoảng thời gian đó. Nếu khoảng thời gian này ngắn thì biểu cảm khuôn mặt cũng xuất hiện trong khoảng thời gian ngắn; và khi đó chỉ có phần liền nét trong mẫu thời gian xuất hiện. Ngược lại, nếu khoảng thời gian này dài, biểu cảm khuôn mặt cũng sẽ xuất hiện trong khoảng thời gian dài. Khi đó ngoài phần liền nét, trong mẫu theo thời gian phần đứt nét cũng có mặt. Như được chỉ ra trong mẫu, mặc dù trạng thái cảm xúc có thể giữ ở cường độ đủ lớn trong khoảng thời gian dài, nhưng trong khoảng thời gian này, biểu cảm khuôn mặt tương ứng không giữ nguyên ở một cường độ. Trái lại, biểu cảm khuôn mặt xuất hiện với cường độ tương ứng với cường độ của trạng thái cảm xúc, sau đó nó giữ nguyên ở trạng thái này trong một khoảng thời gian, và tiếp theo sẽ giảm gần về trạng thái ban đầu. Chúng tôi gọi quá trình này là một chu kỳ. Với cảm xúc buồn và cảm xúc vui, chu kỳ này sẽ lặp một số lần với cường độ giảm dần, sau đó biểu cảm khuôn mặt được giữ ở cường độ thấp cho tới khi kết thúc khoảng thời gian mà trạng thái cảm xúc tồn tại. Với bốn cảm xúc còn lại, biểu cảm khuôn mặt thường chỉ xuất hiện một chu kỳ và sau đó biểu cảm



Hình 3.8: (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 các cảm xúc sợ, giận, ngạc nhiên, và khinh bỉ.

được giữ ở cường độ thấp.

Một chu kỳ biểu cảm được định nghĩa như sau:

$$E = (P, Ts, Te, Do, Dr), \tag{3.4}$$

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

$$F_e(t) = \begin{cases} P.\phi_+(t - Ts, Do) & \text{nếu } (Ts < t < Ts + Do) \\ P & \text{nếu } (Ts + Do \leq t \leq Te - Dr) \\ P.\phi_-(t - Te + Dr, Dr) & \text{nếu } (Te - Dr < t < Te) \end{cases} \tag{3.5}$$

trong đó ϕ_+ và ϕ_- là các hàm mô tả giai đoạn onset và offset của chu kỳ biểu cảm. Theo nghiên cứu của Essa's [44], đường cong hàm mũ được sử dụng để khớp (fit) các phần onset và offset của chu kỳ biểu cảm. Hàm mũ dạng $(e^{bx} - 1)$ được đề xuất cho phần onset, trong khi hàm mũ dạng $(e^{c-dx} - 1)$ được đề nghị cho phần offset. Dựa trên các hàm được gợi ý, chúng tôi có được hai hàm cho

phần onset và phần offset; với phần onset, b được chọn sao cho:

$$\phi_+(0, Do) = e^{b.0} - 1 = 0, \tag{3.6}$$

$$\text{và } \phi_+(Do, Do) = e^{b.Do} - 1 = 1 \tag{3.7}$$

Từ phương trình thứ hai có được hàm mô tả phần onset, hàm này được định nghĩa như sau:

$$\phi_+(x, Do) = exp(\frac{ln2}{Do}x) - 1. \tag{3.8}$$

Với phần offset, c và d được chọn sao cho:

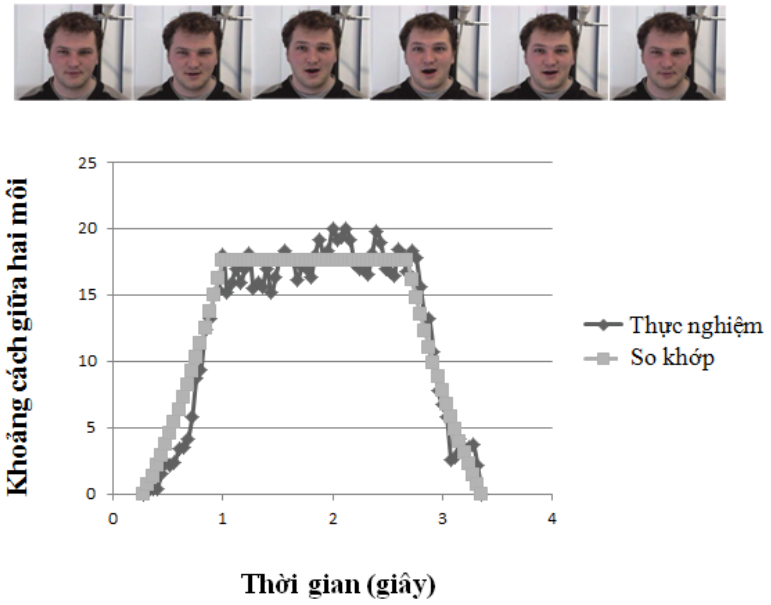
$$\phi_-(0, Dr) = e^{c-d.0} - 1 = 1, \tag{3.9}$$

$$\text{và } \phi_-(Dr, Dr) = e^{c-d.Dr} - 1 = \frac{a}{P} \tag{3.10}$$

Từ hai phương trình trên có được hàm mô tả phần offset, được định nghĩa như sau:

$$\phi_-(x, Dr) = exp(ln2 - \frac{ln2 - ln(\frac{a}{P} + 1)}{Dr}x) - 1. \tag{3.11}$$

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 mẫu AU theo thời gian (temporal AU profiles). Hình 3.9 chỉ ra một ví dụ của mẫu theo thời gian AU25 của một đối tượng thể hiện cảm xúc ngạc nhiên. Người này thể hiện sự tăng dần trong khoảng cách giữa hai môi (Lips Part (AU25)), điển hình cho biểu cảm ngạc nhiên. Trong hình 3.9, phần trên là các frame tại một số thời điểm của video; phần dưới là đồ thị theo thời gian thể hiện khoảng cách giữa hai môi, khoảng cách này đặc trưng cho cường độ của AU25. Trong đồ thị, đường và các điểm tối màu hơn thể hiện dữ liệu thu được từ phân tích thực nghiệm. Đường và các điểm sáng màu hơn thể hiện dữ liệu so khớp (fitting data) khi sử dụng mẫu và hàm được mô tả ở trên. Nếu khoảng cách giữa hai con ngươi mắt được chuẩn hóa bằng 1 thì tổng bình phương lỗi (sum of squares due to error) (SSE) của phép so khớp là 0.0207. Thực hiện thao tác so khớp cho tất cả các mẫu AU theo thời gian chúng tôi thu được giá trị trung bình của tổng bình phương lỗi là 0.055 với độ lệch chuẩn là 0.078. Giá trị này cho thấy mẫu theo thời gian và hàm so khớp ở trên là hợp lý.

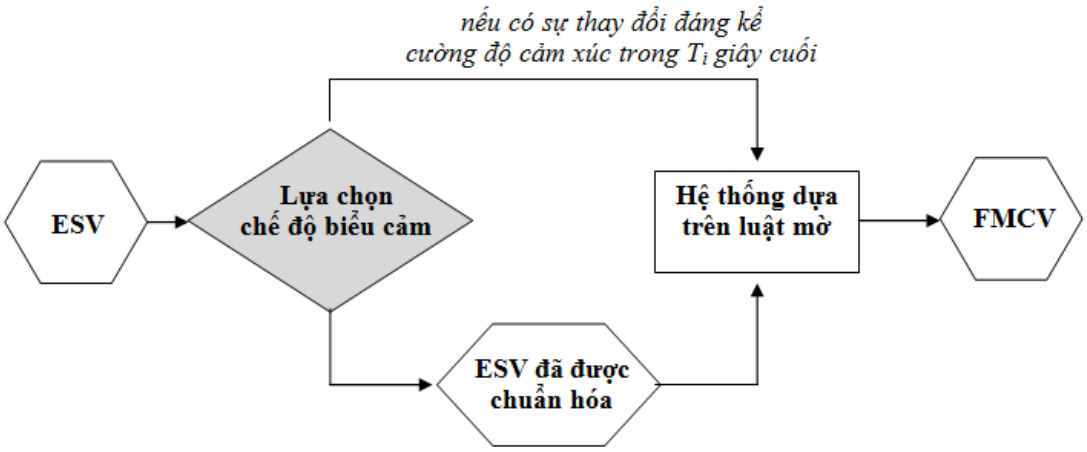


Hình 3.9: Mẫu thực nghiệm và mẫu so khớp theo theo thời gian của AU25 của một người với cảm xúc ngạc nhiên. TRÊN: Video frame được ghi lại tại một số thời điểm từ video tổng thể. DƯỚI: Mẫu theo thời gian thể hiện khoảng cách giữa hai môi, khoảng cách này đặc trưng cho cường độ của AU25.

Kết quả phân tích cho thấy với cảm xúc vui thì khoảng thời gian trung bình của một chu kỳ biểu cảm là 3.5 giây. Khoảng thời gian này thường không nhỏ hơn 1.5 giây và không lớn hơn 6 giây. Với cảm xúc buồn, thời gian trung bình của một chu kỳ biểu cảm là 5.3 giây, khoảng thời gian này thường không nhỏ hơn 2 giây và không lớn hơn 7 giây. Kết quả phân tích cũng cho thấy khoảng thời gian trung bình của một chu kỳ biểu cảm 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. Kết quả phân tích này hoàn toàn phù hợp với kết quả nghiên cứu tâm lý và sinh lý học đã được công bố [56] đó là biểu cảm khuôn mặt thể hiện cảm xúc thường chỉ xuất hiện trong vài giây.

3.3.2.2 Mô hình đề xuất

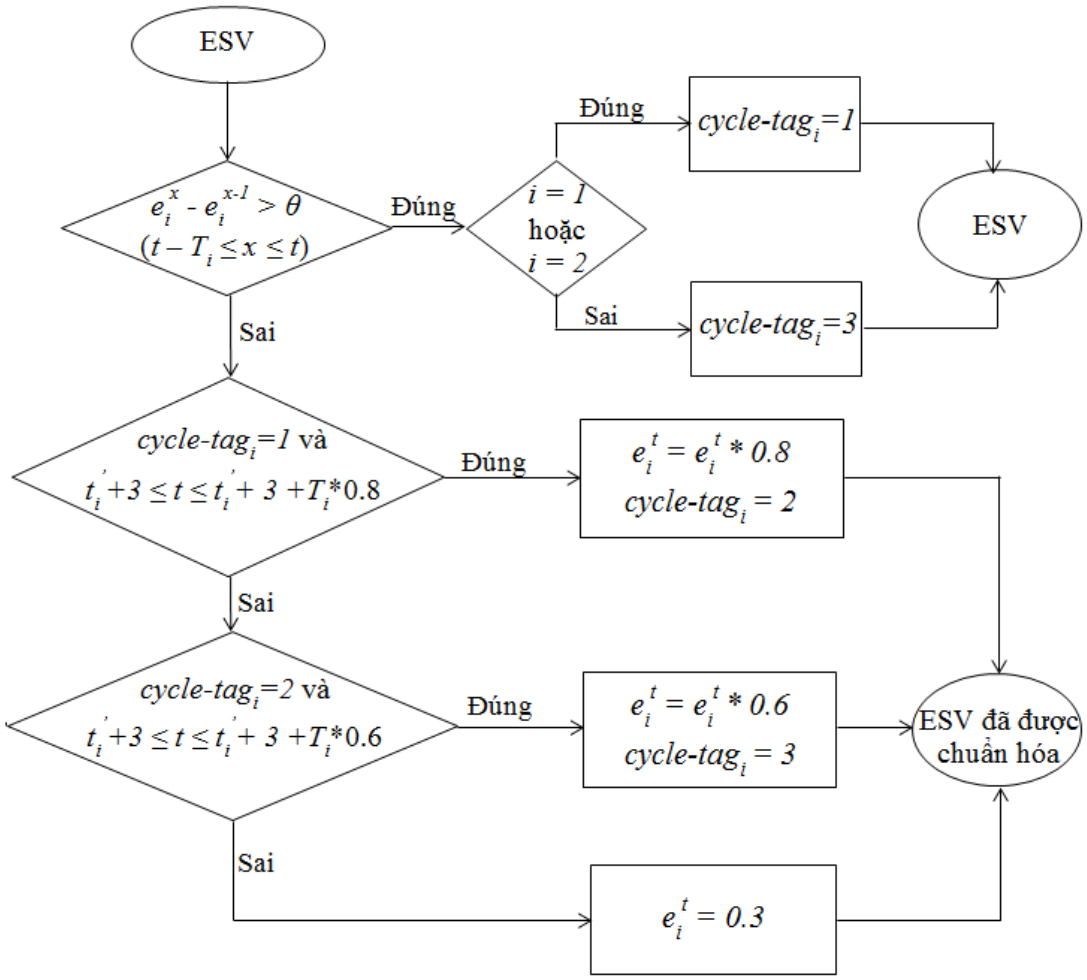
Trong phần này, luận án đề xuất mô hình thứ hai để chuyển trạng thái cảm xúc liên tục của nhân vật ảo thành biểu cảm khuôn mặt. Các mẫu theo thời



Hình 3.10: Mô hình thứ hai chuyển trạng thái cảm xúc liên tục thành biểu cảm khuôn mặt.

gian của cử động khuôn mặt thể hiện các cảm xúc cơ bản (trong phần 3.3.2.1) 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. Ví dụ, khi một sự kiện xảy ra, kích hoạt cảm xúc vui ở một người trong khoảng thời gian khá dài, anh ta/cô ta sẽ không cười với cường độ lớn trong suốt thời gian mà cảm xúc vui tồn tại. Thay vào đó, anh ấy/cô ấy thể hiện một chuỗi biểu cảm cười với cường độ giảm dần. Như vậy, biểu cảm khuôn mặt thể hiện cảm xúc chỉ xuất hiện khi có một tác nhân kích thích làm thay đổi đáng kể trạng thái cảm xúc. Khi không có sự thay đổi đáng kể trong cường độ của các cảm xúc, 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. Biểu cảm thể hiện cảm xúc sẽ không được giữ trên khuôn mặt trong một khoảng thời gian dài, trong khi đó cảm xúc lại triệt tiêu chậm. Tuy nhiên, các biểu cảm của tâm trạng có thể được giữ trên khuôn mặt trong khoảng thời gian dài hơn rất nhiều.

Tương tự như mô hình thứ nhất, mô hình thứ hai cũng dùng hệ thống trong [18] để chuyển một trạng thái cảm xúc tĩnh thành biểu cảm khuôn mặt. Như chỉ ra trên Hình 3.10, mô hình thứ hai cũng gồm bốn thành phần như mô hình thứ nhất. Tuy nhiên, trong mô hình thứ hai, hoạt động của mô đun *Lựa chọn chế độ biểu cảm* được thay đổi so với hoạt động của mô đun này trong mô



Hình 3.11: Hoạt động của mô đun *Lựa chọn chế độ biểu cảm* trong mô hình đề xuất thứ hai.

hình thứ nhất.

Hoạt động của mô đun *Lựa chọn chế độ biểu cảm* được minh họa bởi sơ đồ khối trong Hình 3.11. Cụ thể như sau: Mô đun *Lựa chọn chế độ biểu cảm* điều chỉnh giá trị của chuỗi EVS theo thời gian để biểu cảm tương ứng trên khuôn mặt xuất hiện theo cách tương tự như các mẫu tìm thấy ở phần 3.3.2.1. Mô đun này quyết định việc một biểu cảm thể hiện cảm xúc trên khuôn mặt có được tạo ra để thể hiện trạng thái cảm xúc hiện tại, hay biểu cảm trên khuôn được giữ ở cường độ thấp để thể hiện tâm trạng thay vì cảm xúc. 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 (khoảng thời gian của một chu kỳ biểu cảm thể hiện cảm xúc), tức là nếu:

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

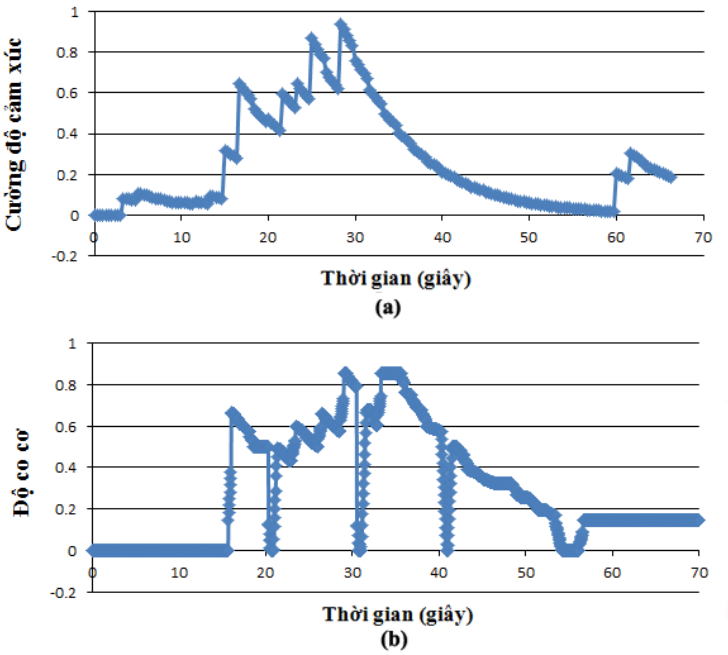
trong đó $t - T_i \leq x \leq t$, t là thời điểm hiện tại, và θ là ngưỡng để kích hoạt các biểu cảm thể hiện cảm xúc trên khuôn mặt (ngưỡng θ được chọn giá trị 0.3 qua thực nghiệm). (Theo kết quả phân tích trong phần 3.3.2.1, T_i có giá trị khoảng 3.5 cho cảm xúc vui, 5.3 cho cảm xúc buồn, 3.6 cho cảm xúc khinh bỉ, 3 cho cảm xúc giận dữ và sợ hãi, và 2.7 cho cảm xúc ngạc nhiên). 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 [18]; 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 (sợ hãi, giận dữ, ngạc nhiên, khinh bỉ). 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 $e_i^t = e_i^t * 0.8$ 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 $e_i^t = e_i^t * 0.6$ và $cycle - tag_i = 3$.
- trường hợp còn lại thì e_i^t được chuẩn hóa về cường độ thấp hơn. Trong trường hợp này, cảm xúc được thể hiện dưới dạng tâm trạng - trạng thái cảm xúc ở cường độ thấp và với thời gian kéo dài.

Sau khi đã được chuẩn hóa, véc tơ EVS được chuyển thành véc tơ FMCV cũng dùng *Hệ thống dựa trên luật mờ* như trong [18].

3.4 Thực nghiệm và đánh giá

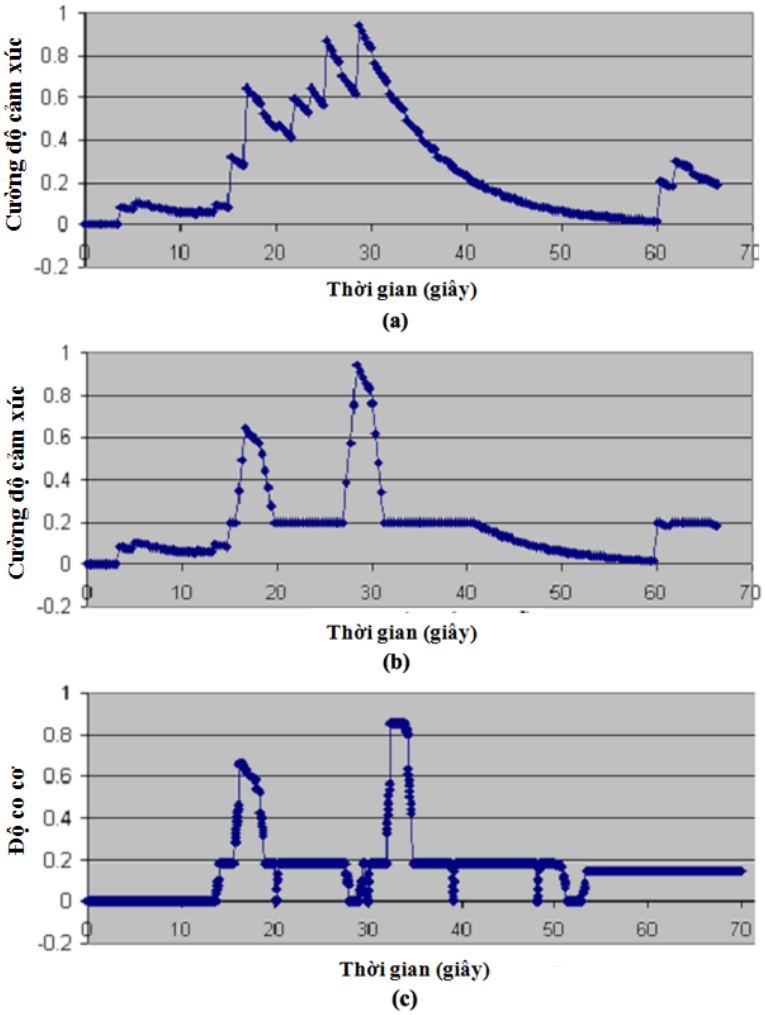
Luận án sử dụng nhân vật ảo trong [16] để đánh giá các mô hình tạo biểu cảm khuôn mặt thể hiện trạng thái cảm xúc liên tục được đề xuất trong Phần 3.3. Theo hiểu biết của chúng tôi, cho tới nay, đây là nhân vật ảo duy nhất có khả năng ánh xạ trạng thái cảm xúc liên tục thành biểu cảm khuôn mặt tương ứng; và hai mô hình được đề xuất trong phần 3.3 được áp dụng vào nhân vật ảo này. Đây là nhân vật ảo được đặt trong miền của một cổ động viên bóng



Hình 3.12: (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 đề xuất.

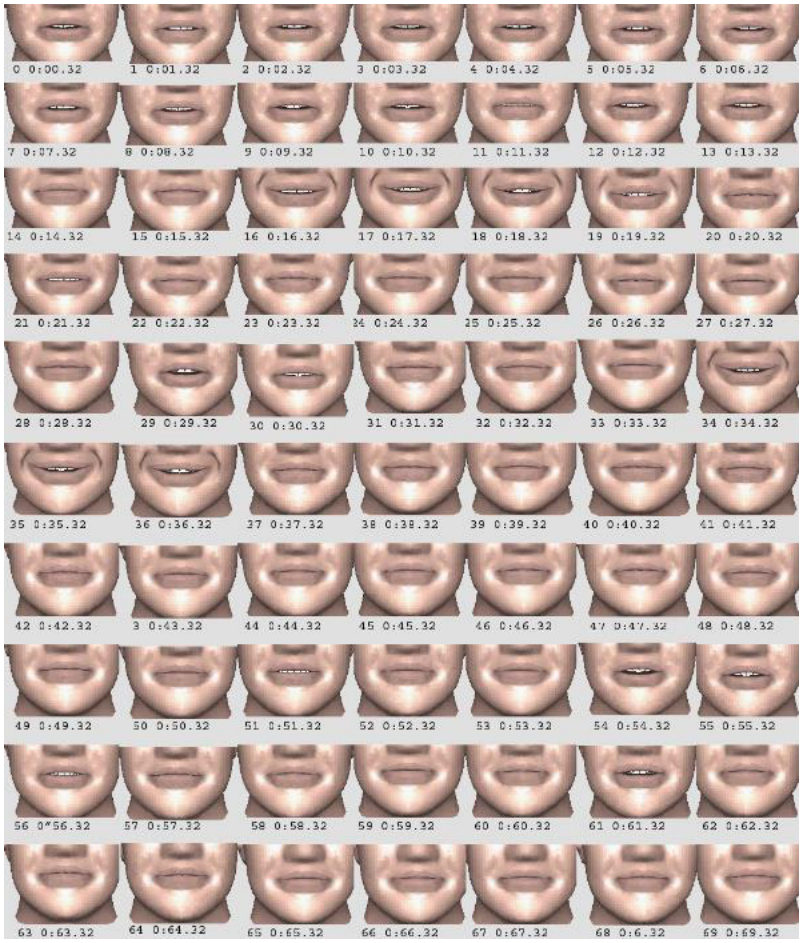
đá. Bóng đá là một trò chơi cảm xúc; có rất nhiều sự kiện trong trò chơi này có thể kích hoạt cảm xúc không chỉ của người chơi mà cả huấn luyện viên và cổ động viên... Bàn thắng ở phút cuối khiến một số người vui mừng, tin tưởng, trong khi đó lại khiến những người khác buồn chán, giận dữ, hay thất vọng. Kiểm tra các mô hình tạo biểu cảm khuôn mặt trong miền của cổ động viên bóng đá cho chúng ta cơ hội để kiểm tra nhiều cảm xúc cũng như tính động của các cảm xúc, vì các hành động, sự kiện trong một trận bóng đá xảy ra khá nhanh. Nhân vật ảo có tên Obie, được mô phỏng như một cổ động viên bóng đá. Obie xem một trận bóng đá trong đó đội anh ta cổ vũ đang chơi. Obie có thể trải nghiệm những cảm xúc khác nhau thông qua việc đánh giá, thẩm định các sự kiện dựa trên mục tiêu, tiêu chuẩn, và sở thích của anh ta. Obie còn có khả năng thể hiện cảm xúc của mình trên khuôn mặt ba chiều; nhân vật ảo này sử dụng cơ chế ánh xạ trực tiếp từ trạng thái cảm xúc thành biểu cảm khuôn mặt.

Trước khi áp dụng mô hình đề xuất để chuyển trạng thái cảm xúc liên tục thành biểu cảm khuôn mặt, thỉnh thoảng Obie thể hiện biểu cảm khuôn mặt với cường độ lớn trong một khoảng thời gian dài. Điều này khiến nhân vật ảo



Hình 3.13: (a): Đồ thị thể hiện cường độ cảm xúc vui của Obie trong trận bóng đá. (b): Đồ thị thể hiện cường độ 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.

có một diện mạo máy móc, không được tự nhiên bởi vì thông thường một biểu cảm thể hiện cảm xúc với cường độ cao chỉ lưu lại trên khuôn mặt con người trong khoảng vài giây. Có thể dễ dàng nhận thấy điều này ở đồ thị mô tả cường độ cảm xúc vui và mức co của cơ Zymgomatic Major - cơ cười thể hiện cảm xúc vui trong Hình 3.12. Hình 3.12(a) thể hiện cường độ cảm xúc vui của Obie trong trận bóng đá và Hình 3.12(b) thể hiện mức co của cơ Zymgomatic Major - cơ cười để diễn tả cảm xúc vui trước khi áp dụng mô hình đề xuất. Từ hình vẽ, chúng ta có thể thấy mức co của cơ Zymgomatic Major giữ ở trạng thái cao trong một khoảng thời gian dài, từ giây 16 tới giây 46. Điều này có nghĩa là

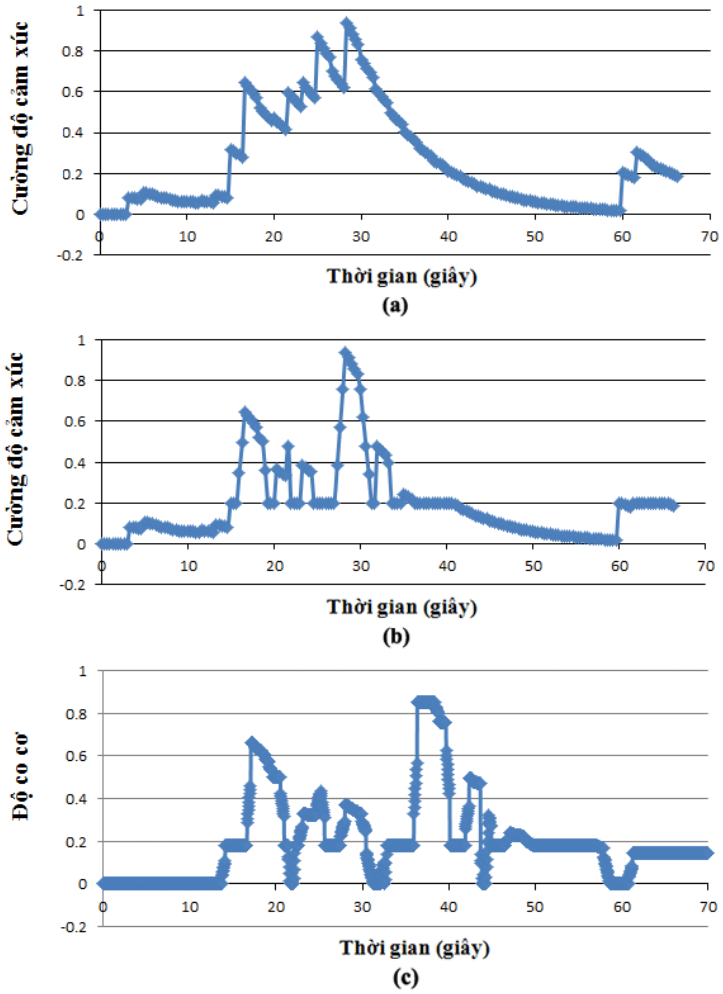


Hình 3.14: Biểu cảm khuôn mặt thể hiện cảm xúc vui trên khuôn mặt ba chiều sau khi áp dụng mô hình đề xuất thứ nhất (lật lượt theo frame).

khuôn mặt của Obie "cười" liên tục trong khoảng 30 giây.

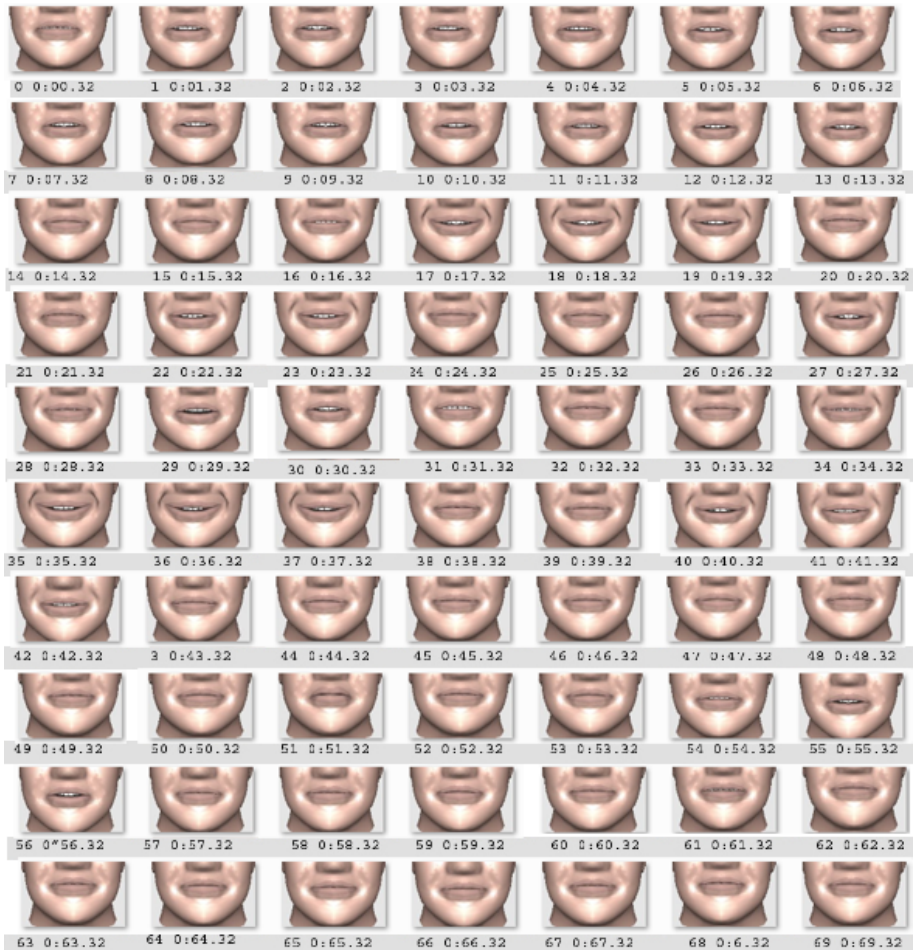
Sau khi áp dụng mô hình đề xuất thứ nhất, Obie thể hiện cảm xúc của anh ta trên khuôn mặt ba chiều theo một cách thức hợp lý hơn. Mỗi biểu cảm trên khuôn mặt chỉ kéo dài khoảng ba giây. Trong khoảng thời gian còn lại, mặc dù cảm xúc vui vẫn còn tồn tại với cường độ cao, khuôn mặt chỉ thể hiện biểu cảm ở cường độ thấp để diễn tả tâm trạng vui. Có thể nhận thấy điều này từ Hình 3.13 và Hình 3.14. Từ các hình vẽ, chúng ta có thể thấy khuôn mặt 3D chỉ "cười" hai lần, tại giây thứ 16 và giây thứ 34, mỗi lần khoảng ba giây. Trong khoảng thời gian còn lại, Obie thể hiện tâm trạng vui mặc dù lúc này cảm xúc vui thực tế vẫn tồn tại với cường độ cao.

Sau khi áp dụng mô hình đề xuất thứ hai, Obie cũng thể hiện cảm xúc của anh ta trên khuôn mặt ba chiều theo một cách thức hợp lý hơn so với cơ chế



Hình 3.15: (a): Đồ thị thể hiện cường độ cảm xúc vui của Obie trong trận bóng đá. (b): Đồ thị thể hiện cường độ 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.

ánh xạ trực tiếp. Khi cảm xúc vui với cường độ cao xảy ra trong khoảng thời gian dài, biểu cảm khuôn mặt tương ứng 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 3.15 và Hình 3.16. Từ các hình vẽ, chúng ta có thể thấy khuôn mặt ba chiều chỉ "cười" hai lần, tại giây thứ 16 và giây thứ 34; mỗi lần có hai hoặc ba chu kỳ biểu cảm, chu kỳ đầu tiên kéo dài 3.5 giây, các chu kỳ sau có khoảng thời gian giảm dần. Trong khoảng thời gian còn lại, Obie thể hiện tâm trạng vui mặc dù lúc này cảm xúc vui thực tế vẫn tồn tại với cường độ cao.



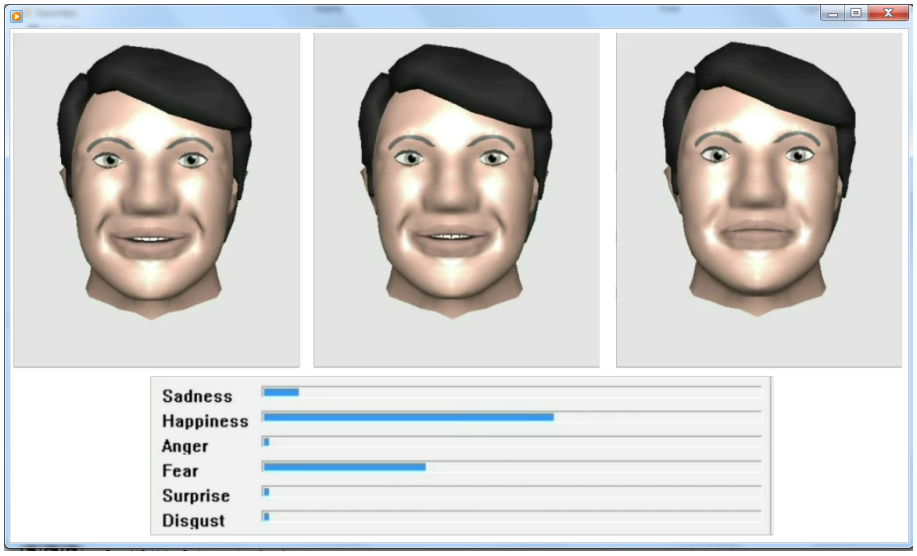
Hình 3.16: Biểu cảm khuôn mặt thể hiện cảm xúc vui trên khuôn mặt ba chiều (lật lượt theo frame) sau khi áp dụng mô hình đề xuất thứ hai.

Thực nghiệm đánh giá với người dùng:

Để đánh giá khả năng tạo biểu cảm khuôn mặt thể hiện trạng thái cảm xúc liên tục của các mô hình được đề xuất, chúng tôi đã tiến hành thực nghiệm để thu thập đánh giá của người dùng. Theo Katherine Isbister và Patrick Doley [67] chúng tôi đã chọn phương pháp đánh giá với người dùng cho các thực nghiệm đánh giá liên quan tới cảm xúc và biểu cảm khuôn mặt. Quá trình tiến hành thực nghiệm và kết quả đánh giá như sau:

Đối tượng được đánh giá: Nhằm đánh giá ưu điểm và hiệu quả của các mô hình đề xuất, 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, được trình bày trong nghiên cứu của tác giả Bui và cộng sự [16]. Như đã đề cập, theo hiểu biết của chúng tôi, cho tới nay, đây là nhân vật ảo duy nhất có



Hình 3.17: Hình ảnh minh họa video clip dùng để đánh giá các mô hình tạo biểu cảm khuôn mặt.

khả năng ánh xạ trạng thái cảm xúc liên tục thành biểu cảm khuôn mặt tương ứng; nó 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 hai.

Chuẩn bị cho thực nghiệm đánh giá:

Để tiến hành thực nghiệm đánh giá, 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, người tham gia thực nghiệm đánh giá không biết trước thứ tự này; 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. Hình ảnh của video clip được minh họa trong Hình 3.17.

Mục tiêu của thực nghiệm là đánh giá tính thuyết phục của các nhân vật ảo A, B, C trong việc tạo biểu cảm khuôn mặt thể hiện trạng thái cảm xúc được chỉ ra ở phần dưới trong hình ảnh của video clip. Những người tham gia thực nghiệm sẽ đánh giá xem các nhân vật ảo có thể hiện đúng trạng thái cảm xúc

Tính thuyết phục của nhân vật ảo bên trái?						
0 (Rất kém)	**1** (Kém)	**2** (Hơi kém)	**3** (Trung bình)	**4** (Khá tốt)	**5** (Tốt)	**6** (Rất tốt)
Tính thuyết phục của nhân vật ảo ở giữa?						
0 (Rất kém)	**1** (Kém)	**2** (Hơi kém)	**3** (Trung bình)	**4** (Khá tốt)	**5** (Tốt)	**6** (Rất tốt)
Tính thuyết phục của nhân vật ảo bên phải?						
0 (Rất kém)	**1** (Kém)	**2** (Hơi kém)	**3** (Trung bình)	**4** (Khá tốt)	**5** (Tốt)	**6** (Rất tốt)

Hình 3.18: Mẫu ghi kết quả đánh giá tính thuyết phục trong việc thể hiện cảm xúc trên khuôn mặt của các nhân vật ảo .

được chỉ ra hay không, cách thể hiện cảm xúc có tự nhiên và hợp lý không.

Tiếp đến, phương pháp ghi lại kết quả đánh giá của người dùng được xây dựng. 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, theo mẫu được chỉ ra trong Hình 3.18.

Kịch bản tiến hành thực nghiệm:

Thực nghiệm đánh giá tính thuyết phục của các nhân vật ảo A, B, C trong việc tạo biểu cảm khuôn mặt thể hiện cảm xúc được tiến hành với 14 người tham gia trong độ tuổi từ 20 đến 35, có khả năng nghe, nói, nhìn bình thường. Thực nghiệm được tiến hành trong phòng kín, cách âm tương đối tốt nhằm hạn chế tối đa ảnh hưởng của các yếu tố bên ngoài đối với kết quả đánh giá của người dùng. Mỗi phiên thực nghiệm được tiến hành riêng cho từng người như sau: Trước tiên người dùng được giới thiệu về video clip, về mục tiêu của thực nghiệm, về mẫu ghi kết quả đánh giá. Tiếp đến, người dùng sẽ xem video clip về ba nhân vật ảo đã đề cập ở trên; số lần xem video clip không bị giới hạn, người dùng có thể yêu cầu xem lại nếu họ muốn. Sau khi đã xem video clip, người dùng được yêu cầu ghi kết quả đánh giá vào mẫu như Hình 3.18. Với mỗi nhân vật ảo, người dùng sẽ khoanh tròn vào số điểm họ chọn.

Kết quả đánh giá:

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**

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

Sau khi tiến hành thực nghiệm, kết quả đánh giá của người dùng được tổng kết trong Bảng 3.3, Hình 3.19, Hình 3.20, và Hình 3.21. 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.

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:

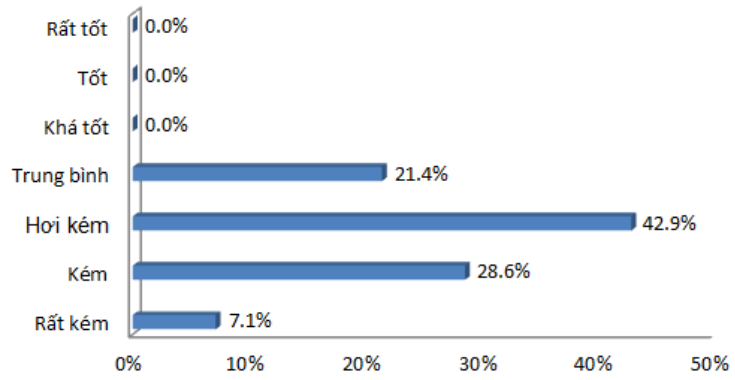
$$H_0 : \mu_A - \mu_B \geq 0,$$

$$H_1 : \mu_A - \mu_B < 0$$

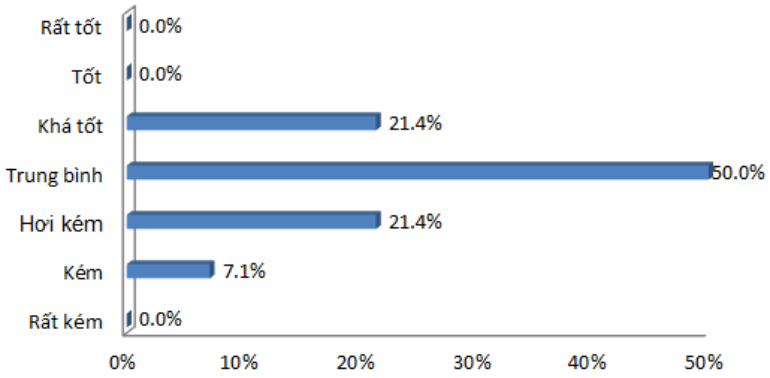
Chúng tôi chọn mức ý nghĩa là 0.05 và sử dụng phương pháp kiểm định *matched-pairs t-test*.

Đặt $D = A - B$, $D_i = A_i - B_i$, D_i nhận các giá trị sau:

$$D_i : -2 \quad -2 \quad -2 \quad -1 \quad -1 \quad -3 \quad 0 \quad -1 \quad 0 \quad -1 \quad 0 \quad 1 \quad -1 \quad -2$$



Hình 3.19: Kết quả đánh giá tính thuyết phục trong việc tạo biểu cảm khuôn mặt của nhân vật ảo A.



Hình 3.20: Kết quả đánh giá tính thuyết phục trong việc tạo biểu cảm khuôn mặt của nhân vật ảo B.

Ta có $\bar{d} = -1.07143$.

Từ đó, độ lệch chuẩn của sự khác nhau được tính từ các cặp đôi là:

$$s_d = sqrt[\Sigma(d_i - \overline{d})^2/(n-1)] = 1.07161.$$

Lỗi chuẩn của phân phối lấy mẫu của $\overline{d}$ là

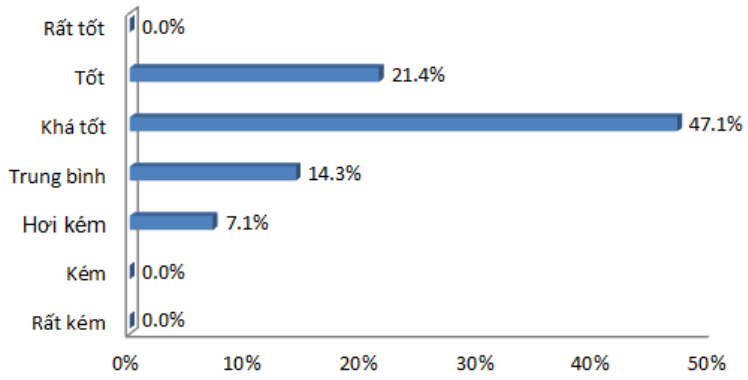
$$SE = s_d/sqrt(n) = 0.28640.$$

Độ tự do $DF = n - 1 = 13$.

Từ đó tính được $t = [(\overline{x_1} - \overline{x_2}) - D]/SE = (\overline{d} - D)/SE = -3.74102$.

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; trung bình điểm đánh giá tính thuyết phục của nhân vật ảo B (2.857) lớn hơn về mặt thống kê so với



Hình 3.21: Kết quả đánh giá tính thuyết phục trong việc tạo biểu cảm khuôn mặt của nhân vật ảo C.

trung bình điểm đánh giá tính thuyết phục của nhân vật ảo A (1.786). Từ kết quả này, 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:

$$H_0 : \mu_B - \mu_C \geq 0,$$

$$H_1 : \mu_B - \mu_C < 0$$

Tương tự như trên, chúng tôi chọn mức ý nghĩa là 0.05 và sử dụng phương pháp kiểm định *matched-pairs t-test*.

Đặt $D = B - C$, $D_i = B_i - C_i$, D_i nhận các giá trị sau:

$$D_i : -1 \quad -1 \quad 0 \quad -2 \quad -1 \quad -1 \quad -1 \quad -1 \quad -1 \quad -1 \quad -1 \quad -1 \quad -1 \quad -2$$

Ta có $\bar{d} = -1.07143$.

Từ đó, độ lệch chuẩn của sự khác nhau được tính từ các cặp đôi là:

$$s_d = sqrt[(\Sigma(d_i - \bar{d})^2/(n-1)] = 0.47463.$$

Lỗi chuẩn của phân phối lấy mẫu của $\bar{d}$ là

$$SE = s_d/sqrt(n) = 0.12685.$$

Độ tự do $DF = n - 1 = 13$.

Từ đó tính được $t = [(\overline{x_1} - \overline{x_2}) - D]/SE = (\bar{d} - D)/SE = -8.44639$.

Từ giá trị t ở trên, ta có $P = 0.00000$.

Vì $P = 0.00000 < 0.05$ nên giả thuyết H_0 bị từ chối; trung bình điểm đánh giá tính thuyết phục của nhân vật ảo C (3.929) lớn hơn về mặt thống kê so với trung bình điểm đánh giá tính thuyết phục của nhân vật ảo B (2.857). Từ kết quả này, 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.

Như vậy, nhân vật ảo B thuyết phục hơn nhân vật ảo A trong việc tạo biểu cảm thể hiện cảm xúc trên khuôn mặt; và nhân vật ảo C thuyết phục hơn nhân vật ảo B trong việc tạo biểu cảm thể hiện cảm xúc trên khuôn mặt. 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

Chương 3 của 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. Ý tưởng của mô hình thứ nhất là một biểu cảm khuôn mặt xuất hiện trong vài giây chỉ khi có sự thay đổi đáng kể của trạng thái cảm xúc. Ý tưởng này xuất phát từ kết quả nghiên cứu tâm lý và sinh lý học rằng một biểu cảm khuôn mặt thường chỉ xuất hiện trong vài giây. Ý tưởng của mô hình thứ hai 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. Ý tưởng này xuất phát từ kết quả của 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. Các thực nghiệm đánh giá đã được thực hiện, và kết quả 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. Vì vậy, 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

Tiếng nói là một trong những phương thức thuận tiện và quan trọng nhất mà con người sử dụng để giao tiếp với nhau. Rõ ràng chúng ta không chỉ dùng thông tin ngôn ngữ để truyền tải ý định, cảm giác mà chúng ta còn vô tình hay hữu ý đưa cảm xúc của chúng ta vào tiếng nói. Như đã đề cập trong Chương 3, Mehrabian [98] đã chỉ ra rằng trong giao tiếp trực tiếp người - người, chỉ có 7% thông điệp cảm xúc được truyền tải qua từ ngữ, trong khi đó có tới 38% thông điệp được truyền tải qua yếu tố giọng điệu. Mối quan hệ giữa cảm xúc và tiếng nói đã được tổng kết trong Chương 2; các nghiên cứu đã chỉ ra rằng tồn tại mối liên hệ giữa trạng thái cảm xúc và giọng điệu khi phát âm. Cảm xúc đóng vai trò cực kỳ quan trọng trong suốt quá trình giao tiếp của con người. Vì lý do này, các nhà nghiên cứu đã và đang cố gắng đưa cảm xúc vào thế giới ảo nhằm tăng cường tính tự nhiên của chúng. Và nhằm mục đích khiến cho giao diện tương tác của các hệ thống hội thoại giống với con người hơn, việc cố gắng đưa cảm xúc vào tiếng nói tổng hợp là cần thiết. Từ đó có thể thấy, với bài toán thể hiện cảm xúc cho nhân vật ảo thì ngoài khuôn mặt, tiếng nói cũng là một kênh biểu cảm quan trọng cần được quan tâm.

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 không cảm xúc. Trước tiên, chúng tôi mô tả một số phân tích về đặc trưng âm của tiếng nói tiếng việt có cảm xúc. Việc phân

tích được thực hiện nhằm tìm ra mối quan hệ giữa sự biến đổi của ngôn điệu, âm sắc với trạng thái cảm xúc trong tiếng nói tiếng Việt. Cụ thể, một cơ sở dữ liệu tiếng nói tiếng Việt có cảm xúc, đa trạng thái, được xây dựng và phân tích nhằm xác minh mối tương quan và định lượng cho các trạng thái cảm xúc về sự biến đổi của các đặc trưng ngôn điệu và đặc trưng âm sắc so với trạng không cảm xúc. Dựa trên kết quả phân tích, tập các hệ số biến đổi ngôn điệu và âm sắc được đưa ra cho mỗi trạng thái cảm xúc. Sau đó, tần số cơ bản đích cùng với các ràng buộc về thời gian, năng lượng, phổ, được tạo ra bằng cách áp dụng các luật suy ra từ tập hệ số nói trên. Quá trình phân tích cơ sở dữ liệu được thực hiện ở mức phát âm toàn câu và mức âm tiết; và các luật được suy ra có tính đến sự biến đổi đặc trưng âm ở mức âm tiết. Từ đó, tiếng nói ở trạng thái không cảm xúc được biến đổi để tạo ra tiếng nói tổng hợp có cảm xúc. Trong quá trình tổng hợp tiếng nói có cảm xúc này, đặc trưng âm được biến đổi nhiều hơn ở một số âm tiết thay vì biến đổi đồng đều trong tất cả các âm tiết của câu. Đây là điểm khác so với các nghiên đã được đề xuất, khiến cho tiếng nói tổng hợp có cảm xúc tự nhiên hơn, thực hơn.

Nội dung của chương được tổ chức như sau. Phần 4.2 trình bày tóm tắt về các nghiên cứu liên quan. Tiếp theo, Phần 4.3 mô tả giai đoạn trích đặc trưng âm liên quan tới tiếng nói tiếng Việt có cảm xúc và kết quả phân tích. Sau đó, Phần 4.4 mô tả việc xây dựng các luật dùng để tổng hợp tiếng nói tiếng Việt có cảm xúc từ tiếng nói không cảm xúc; phần này cũng chỉ ra tiến trình tổng hợp tiếng nói tiếng Việt có cảm xúc. Kết quả đánh giá sẽ được trình bày trong Phần 4.5.

4.2 Những nghiên cứu liên quan

4.2.1 Các phương pháp tổng hợp tiếng nói có cảm xúc

Tổng hợp tiếng nói là quá trình chuyển thông điệp từ chữ viết thành thông điệp tương đương ở dạng tiếng nói. Tổng hợp tiếng nói có cảm xúc bao hàm tổng hợp tiếng nói và thêm vào tiếng nói tổng hợp các biểu cảm khác nhau liên quan tới các cảm xúc khác nhau. Theo các nghiên cứu [131, 53], 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 điểu 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.

Trong loại đầu tiên - tổng hợp tiếng nói có cảm xúc bằng điều khiển tường minh, tiếng nói có cảm xúc được tổng hợp thông qua việc biến đổi tiếng nói không cảm xúc dựa trên một số luật thu được từ cơ sở dữ liệu tiếng nói cảm xúc. Các hệ thống tổng hợp tiếng nói có cảm xúc được phát triển từ phương pháp tổng hợp formant [22, 104, 20] hay phương pháp kết nối diphone [149, 103] là những ví dụ của tổng hợp tiếng nói có cảm xúc bằng điều khiển tường minh. Bên cạnh đó, các phương pháp được đưa ra cho thao tác chuyển tiếng nói từ trạng thái không cảm xúc sang trạng thái có cảm xúc như [139, 125, 64, 141, 21, 61] cũng thuộc vào loại tổng hợp bằng điều khiển tường minh. Kỹ thuật biển đổi tiếng nói này thực hiện một số thao tác trên các tham số đặc trưng âm của dữ liệu tiếng nói để tạo các cảm nhận cảm xúc khác nhau [129]. Kỹ thuật này sử dụng tiếng nói của một câu hoàn chỉnh đã được ghi âm hoặc được tổng hợp từ trước. Thông thường tiếng nói ở trạng thái không cảm xúc được dùng như đầu vào và một số phương pháp biến đổi đặc trưng âm hoặc các kỹ thuật khác được sử dụng để chuyển tiếng nói nguồn thành tiếng nói đích có cảm xúc. Việc sử dụng tiếng nói ở trạng thái không cảm xúc làm đầu vào khiến cho tính tính chân thực (tự nhiên) của tiếng nói tổng hợp được đảm bảo phần nào. Bên cạnh đó, việc sử dụng các kỹ thuật biến đổi tiếng nói có thể cho chúng ta khả năng điều khiển các tham số âm một cách linh hoạt.

Trong tổng hợp tiếng nói có cảm xúc bằng phương pháp phát lại, tiếng nói có cảm xúc được tổng hợp một cách độc lập thông qua việc sử dụng cơ sở dữ liệu tiếng nói cảm xúc tương ứng. Ở phương pháp này, việc tổng hợp tiếng nói có cảm xúc đạt được bằng cách chỉ đơn thuần là phát lại những gì có sẵn trong cơ sở dữ liệu cảm xúc tương ứng, hoặc sử dụng các mô hình được huấn luyện từ cơ sở dữ liệu cảm xúc tương ứng. Các hệ thống tổng hợp tiếng nói có cảm xúc dựa trên lựa chọn đơn vị (unit selection) và HMM, được huấn luyện trên cơ sở dữ liệu cảm xúc tương ứng [155, 63, 60, 45, 116] thuộc phương pháp phát lại. Phương pháp này không có khả năng mô hình hóa một cách tường minh các đặc tính âm của cảm xúc, và nó cũng cần cơ sở dữ liệu tiếng nói rất lớn để có thể cải thiện tính tự nhiên của tiếng nói tổng hợp.

Trong phương pháp điều khiển không tường minh, biểu cảm trong tiếng

nói tổng hợp được điều khiển bởi phép nội suy giữa hai mô hình thống kê được huấn luyện trên các cơ sở dữ liệu cảm xúc khác nhau. Một số nghiên cứu đã công bố sử dụng phép nội suy và cải tiến mô hình HMM để tổng hợp tiếng nói có cảm xúc [101, 154]. Các kỹ thuật cải tiến tạo nên sự linh động trong việc xây dựng các mô hình thống kê với lượng dữ liệu nhỏ nếu như đã có sẵn mô hình trung bình; các kỹ thuật này cũng có thể được sử dụng để tổng hợp tiếng nói ở các trạng thái cảm xúc khác nhau [10, 107]. Tuy nhiên, nhược điểm đáng chú ý của các hệ thống tổng hợp dựa trên HMM là tính tự nhiên của tiếng nói tổng hợp có cảm xúc bị giảm do đặc tính vốn có của các mô hình HMM là quá làm mịn (over-smoothing) các tham số âm.

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

Để có khả năng thực hiện việc đưa cảm xúc vào tiếng nói tổng hợp, về mặt âm học chúng ta cần phải có hiểu biết chi tiết về việc các đặc trưng âm trong tiếng nói liên quan như thế nào đến 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: một loại liên quan tới ngôn điệu và loại còn lại liên quan đến âm sắc.

Ngôn điệu: Ngôn điệu về cơ bản là một tập các yếu tố điều khiển cao độ, độ to, và tốc độ của tiếng nói. Sự biến đổi của âm điệu, nhịp điệu, kiểu nhấn chính là những cái mà chúng ta gọi là ngôn điệu của một câu. Phụ thuộc vào trạng thái cảm xúc của người nói, một câu có thể được phát âm với các đặc tính ngôn điệu khác nhau. Vì vậy, sự biến đổi ngôn điệu trong một phát âm có ảnh hưởng rất lớn đến cảm xúc được thể hiện trong tiếng nói [78]. Đây chính là lý do mà ngôn điệu là một trong những yếu tố quan trọng cần được khảo sát khi tìm sự biến đổi đặc trưng âm liên quan tới trạng thái cảm xúc trong tiếng nói. 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. Những đặc trưng này được mô tả sau đây.

Tần số cơ bản: Đường F0 thể hiện sự thay đổi của F0 trong miền thời gian, nó cung cấp thông tin về điểm nhấn và âm điệu trong phát âm của câu. Những thông tin như vậy có ảnh hưởng rất lớn đối với sự cảm nhận trạng thái cảm xúc trong tiếng nói. Vì vậy, trong lĩnh vực nghiên cứu tiếng nói có cảm xúc, F0 là

đặc trưng âm đã và đang được nghiên cứu thường xuyên và từ thời điểm sớm nhất. Tác giả Erickson [43] đã đưa ra tóm tắt về các nghiên cứu trước đây, nội dung của những nghiên cứu này là tìm ra loại đặc trưng âm nào có liên quan đến trạng thái cảm xúc trong tiếng nói. Hầu hết các nghiên cứu đều nhận thấy rằng đường F0 có ảnh hưởng lớn lên trạng thái cảm xúc trong tiếng nói, bất kể là phương pháp thu thập dữ liệu nào được sử dụng hay ngôn ngữ được dùng là ngôn ngữ gì.

Khoảng thời gian: Về khía cạnh vật lý, khoảng thời gian chủ yếu được nhận thấy như là độ dài, khoảng dừng, và tốc độ của phát âm tiếng nói. Nó thay đổi khá nhiều khi người nói ở trong các trạng thái cảm xúc khác nhau. Trong lĩnh vực nghiên cứu tiếng nói có cảm xúc, đã có các nghiên cứu chỉ ra ảnh hưởng của khoảng thời gian đối với trạng thái cảm xúc trong tiếng nói, ở các ngôn ngữ khác nhau ví dụ tiếng Anh, tiếng Nhật, tiếng Ý [62, 94, 113, 130]. Vì vậy, khoảng thời gian cũng là một trong những yếu tố quan trọng cần được khảo sát.

Năng lượng: Năng lượng của tiếng nói được quyết định bởi thể tích luồng khí của hơi được gửi ra từ phổi, chủ yếu nó thể hiện độ to của âm được cảm nhận bởi người nghe. Tương tự như đường F0, hình bao năng lượng cũng có ảnh hưởng đối với trạng thái cảm xúc trong tiếng nói. Năng lượng có thể biến đổi rộng khi người nói ở các trạng thái cảm xúc khác nhau. Mối quan hệ giữa hình bao năng lượng và trạng thái cảm xúc trong tiếng nói đã được nêu ra trong một số nghiên cứu trước đây (ví dụ [62, 90, 130]). Vì vậy, để nghiên cứu mối quan hệ giữa đặc trưng âm và cảm xúc trong tiếng nói, năng lượng là một yếu tố nữa cần được khảo sát.

Âm sắc: Ngoài các đặc trưng âm liên quan tới ngôn điệu, âm sắc cũng là một trọng điểm chính yếu nữa mà các nhà nghiên cứu đặt rất nhiều sự quan tâm. Một định nghĩa của âm sắc đó là "đặc tính của một âm mà nhờ đó người nghe có thể chỉ ra rằng hai âm với cùng độ to và cao độ nhưng lại không giống nhau" ("the quality of a sound by which a listener can tell that two sounds of the same loudness and pitch are dissimilar" (ANSI, 1973. Psychoacoustical terminology. Technical Report, S.3.30, American National Standard Report). 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. Mặc dù phổ không phải là đặc trưng được khảo sát nhiều như tần số F0, nhưng nó đã được chỉ ra rằng có mối

quan hệ với việc cảm nhận tiếng nói có cảm xúc (ví dụ [90, 94, 130]). 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.

Tần số cộng hưởng (Formants): Khi tiếng nói được tạo ra, sự rung của dây thanh tạo ra các cộng hưởng trong hệ thống phát âm. Các tần số cộng hưởng này thường được xem xét trong phổ của tín hiệu tiếng nói. Về mặt lý thuyết, có một số lượng vô hạn các tần số cộng hưởng, nhưng chỉ có ba hay bốn tần số thấp nhất được quan tâm tới trong các mục đích thực tế. Trong nghiên cứu của Ishii và Campbell [68], các tác giả đã khảo sát sự tương quan giữa đặc trưng âm và một số danh mục của âm sắc và cách thức nói. Kết quả nghiên cứu đã chỉ ra rằng tần số cộng hưởng là một trong những tham số có ảnh hưởng lớn nhất tới âm sắc. Vì vậy, tần số cộng hưởng là một trong những đặc trưng âm có ý nghĩa, cần được khảo sát.

Hình dạng phổ (Spectral shape): Hình dạng phổ có thể cung cấp dấu hiệu có liên quan tới các khía cạnh của âm sắc, ví dụ như H1-A3. Đây là tỉ lệ giữa biên độ của họa ba thứ nhất và biên độ của đỉnh cộng hưởng phổ thứ ba (A3), nó được tác giả Hanson [57] dùng để mô tả độ nghiêng phổ. Như được chỉ ra bởi Manezes và Maekawa [100], H1-A3 chính là các đặc tính chu kỳ thanh môn, ví dụ như tốc độ đóng thanh môn. Nhìn chung, độ nghiêng phổ (H1-A3) liên quan tới "độ sáng" của âm sắc. Phổ với độ nghiêng bằng phẳng hơn, năng lượng trong vùng tần số cao lớn hơn, thể hiện độ sáng của giọng nói; trong khi đó phổ với độ nghiêng rõ ràng hơn, năng lượng trong vùng tần số cao yếu hơn, thể hiện độ tối của giọng nói.

Là ngôn ngữ đơn âm tiết và có thanh điệu, tiếng Việt có những đặc trưng riêng biệt so với tiếng của các ngôn ngữ phương Tây (ngôn ngữ đa âm tiết). 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. Tác giả Lê [89] đã đề xuất và chứng minh năm giả thuyết liên quan đến khoảng khoảng thời gian của tiếng nói tiếng Việt, dựa trên việc phân tích 36 file âm thanh gồm 20.815 từ được đọc bởi các phát thanh viên đến từ các vùng khác nhau của Việt Nam. Theo nghiên cứu [146], các yếu tố có ảnh hưởng tới khoảng thời gian của một đơn vị ngữ âm tiếng Việt là vị trí, cao độ, và cấu trúc của đơn vị ngữ âm đó. Trong nghiên cứu [65], từ ghép tiếng Việt và cấu

trúc cụm từ được khảo sát về mặt tương quan ngữ âm của việc nhấn từ; các đặc trưng âm và đặc trưng cảm giác của các từ ghép tiếng Việt và các mệnh đề đối của chúng đã được đưa ra. Trong nghiên cứu [87, 88], tác giả Le đã mô tả một số kết quả nghiên cứu về đặc trưng âm của tiếng nói tiếng Việt nhằm giúp cho việc tổng hợp tiếng nói tiếng Việt từ văn bản. Tác giả Mac [93] đã đưa ra nghiên cứu về đặc trưng ngôn điệu thính giác-thị giác (audio-visual)của tiếng Việt; nghiên cứu này đã chỉ ra sự đóng góp tương đối của các thông tin thính giác, thị giác, thính giác-thị giác trong việc cảm nhận các đặc trưng ngôn điệu, và chỉ ra người bản xứ và người không phải là bản xứ nhận dạng và nhầm lẫn các đặc trưng này như thế nào. Các tác giả cũng thực hiện phân tích ngôn điệu tiếng nói để xác thực thêm các kết quả của thực nghiệm cảm nhận, và để đưa ra một số đặc trưng ngôn điệu của tiếng Việt. Trong nghiên cứu [151], Vu và cộng sự đã đề xuất phương pháp sử dụng HMM để tổng hợp tiếng nói tiếng Việt. Các nghiên cứu [105, 150] cũng đề xuất các hệ thống tổng hợp tiếng nói tiếng Việt từ text... Tuy nhiên, hầu hết các nghiên cứu đã công bố đều tập trung vào tiếng nói tiếng Việt không cảm xúc; theo hiểu biết của chúng tôi, đến nay chưa có nghiên cứu nào cung cấp khả năng thể hiện cảm xúc trong giọng nói tiếng Việt cho nhân vật ảo.

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

4.3.1 Cơ sở dữ liệu

Xây dựng cơ sở dữ liệu là một bước rất quan trọng trong quá trình tổng hợp tiếng nói có cảm xúc bằng phương pháp điều khiển tường minh. Lý do là vì việc phân tích và đánh giá, ước lượng các đặc trưng âm liên quan tới cảm xúc sẽ được thực hiện trên cơ sở dữ liệu này để xây dựng các luật tường minh dùng cho việc tổng hợp tiếng nói có cảm xúc. Luận án sử dụng cơ sở dữ liệu "diễn" thay vì cơ sở dữ liệu tự nhiên vì những lý do như sau. Trước tiên, để đạt được mục tiêu nghiên cứu, chúng tôi cần có một cơ sở dữ liệu gồm các phát âm của các câu giống nhau nhưng ở các trạng thái cảm xúc khác nhau. Rất khó (và hầu như không thể) có được điều này với cơ sở dữ liệu tự nhiên. Hơn nữa, mục tiêu của luận án là cung cấp tiếng nói có cảm xúc cho các nhân vật ảo như người kể

Bảng 4.1: Kết quả nhận dạng cơ sở dữ liệu tiếng nói có cảm xúc.

		không cảm xúc	vui	hơi giận	buồn	rất giận
Nam	không cảm xúc	**95.37%**	12.3%	3.11%	4.29%	0%
	vui	3.98%	**85%**	0.44%	0%	0%
	hơi giận	0%	2.7%	**89.23%**	1.57%	0%
	buồn	0.65%	0%	0%	**94.14%**	0%
	rất giận	0%	0%	7.22%	0%	**100%**
Nữ	không cảm xúc	**94.44%**	8.24%	2.44%	4.86%	0%
	vui	2.3%	**86.56%**	1.14%	0%	0%
	hơi giận	0%	5.2%	**90.41%**	2.3%	0%
	buồn	3.26%	0%	0%	**92.84%**	0%
	rất giận	0%	0%	6.1%	0%	**100%**

chuyện ảo, cổ động viên bóng đá ảo...; với các nhân vật ảo này, tiếng nói "diễn" có thể chấp nhận được.

Cơ sở dữ liệu tiếng nói có cảm xúc được sử dụng cho việc khảo sát đặc trưng ngôn điệu và đặc trưng âm sắc bao gồm các phát âm tiếng Việt được tạo ra bởi hai nghệ sĩ Việt nổi tiếng, một nghệ sĩ nam và một nghệ sĩ nữ. Hai nghệ sĩ được yêu cầu tạo ra các phát âm ở năm trạng thái khác nhau. Họ phải phát âm 19 câu ở bốn trạng thái cảm xúc cơ bản: vui, buồn, hơi giận, rất giận. Bên cạnh đó, họ cũng tạo ra phát âm của 19 câu này ở trạng thái không cảm xúc. Kết quả là mỗi câu sẽ có một phát âm ở mỗi một trong năm trạng thái, cho cả giọng nam và giọng nữ. Vì vậy, cơ sở dữ liệu sẽ có tổng số 190 phát âm – một nửa trong đó là của giọng nam, nửa còn lại của giọng nữ. Các câu trong cơ sở dữ liệu có độ dài khoảng tám từ và thể hiện tốt bảng ký tự ngữ âm tiếng Việt. Hầu hết chúng đều không chứa nội dung ý nghĩa cảm xúc, vì vậy chúng không thể có ảnh hưởng lên hai nghệ sĩ trong việc sinh ra thái độ cảm xúc đặc biệt. Trong quá trình ghi âm, hai nghệ sĩ lần lượt thể hiện mỗi trạng thái cảm xúc, và luôn có một người quản lý ở đó để giám sát ngôn điệu cũng như phát âm của họ nhằm tránh sự "diễn" thái quá. Tín hiệu âm thanh được ghi lại trong phòng cách âm, sử dụng míc và các thiết bị thu dữ liệu số chất lượng cao; sóng âm được số hóa với tần số 22050 Hz và hệ số lượng tử hóa là 16 bit.

Thực nghiệm kiểm tra cảm nhận của người về cảm xúc trong các phát âm đã được thực hiện với 12 tình nguyện viên, nhằm đánh giá cơ sở dữ liệu và xác

minh xem nó có thể được sử dụng để trích ra các luật liên quan đến các trạng thái cảm xúc không. Mười hai sinh viên có ngôn ngữ mẹ đẻ là tiếng Việt, trong độ tuổi trung bình là 21 tuổi, với khả năng nghe hoàn toàn bình thường, đã tham gia vào quá trình thực nghiệm. Những sinh viên này được yêu cầu đánh giá 190 phát âm tùy theo mức độ cảm nhận của họ về mỗi một trong năm trạng thái cảm xúc. Có tổng số 5 điểm cho mỗi phát âm. Theo đó, nếu một người nhận thấy rằng một phát âm thuộc về một trạng thái cảm xúc mà không có sự băn khoăn hay nhập nhằng nào thì họ sẽ cho trạng thái đó 5 điểm. Ngược lại, nếu người đó phân vân giữa hai hay nhiều hơn hai trạng thái cảm xúc thì họ sẽ chia 5 điểm cho những trạng thái này theo tỉ lệ phù hợp. Các phát âm được trình bày đến người đánh giá theo thứ tự ngẫu nhiên thông qua tai nghe ở mức âm lượng vừa phải, trong một phòng cách âm. Kết quả của thực nghiệm đánh giá được thể hiện trong Bảng 4.1. Mặc dù vẫn có một số nhầm lẫn giữa các trạng thái, nhưng nhìn chung tỉ lệ nhận dạng thu được là cao.

4.3.2 Giai đoạn trích đặc trưng âm

Phần này mô tả giai đoạn trích đặc trưng âm cho tiếng nói tiếng Việt có cảm xúc. Việc quyết định sẽ khảo sát đặc trưng âm nào được thực hiện thông qua xem xét các nghiên cứu đã công bố và dựa trên đặc tính riêng của tiếng Việt. Như đã đề cập trong Phần 4.2.2, nhiều nghiên cứu trước đây đều thống nhất rằng hai loại đặc trưng âm luôn được xem là yếu tố quan trọng nhất liên quan tới trạng thái cảm xúc trong tiếng nói đó là ngôn điệu và âm sắc. Là ngôn ngữ đơn âm tiết và có thanh điệu, tiếng Việt có những đặc trưng riêng biệt so với ngôn ngữ châu Âu (đa âm tiết). Ngôn điệu (prosody) tiếng Việt liên quan tới nhịp điệu giữa các từ trong các nhóm từ hoặc trong các từ ghép, trong khi đó âm điệu (intonation) có ảnh hưởng tổng thể lên toàn bộ câu. Mỗi âm tiết tiếng Việt có thể được xem như là tổ hợp của ba thành phần: phần mở đầu, phần kết thúc, và thanh điệu. Thanh điệu gắn với âm tiết như là bộ phận tổng thể, nó đóng vai trò quan trọng trong toàn bộ âm tiết. Tuy nhiên, đặc trưng thanh điệu không rõ ràng như các đặc trưng khác của tín hiệu tiếng nói. Trong phụ âm đầu tiên chúng ta có thể đã nghe thấy chút ít thanh điệu. Thanh điệu trở nên rõ ràng hơn ở vần và kết thúc hoàn toàn ở cuối âm tiết. Hiện tượng lan tỏa này định rõ bản chất không tuyến tính của thanh điệu. Vì vậy, với ngôn ngữ

đơn âm tiết như tiếng Việt, một âm tiết không dễ dàng được chia thành các phần âm nhỏ như các ngôn ngữ Châu Âu.

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. Đường F0, hình bao năng lượng, và phổ được tính sử dụng STRAIGHT [77] với độ dài FFT là 1024 điểm và tốc độ frame là 1ms. Tần số lấy mẫu là 22050 Hz. Khoảng thời gian được xác định bằng tay với sự hỗ trợ một phần của Wavesurfer [134].

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 được trích ra dùng STRAIGHT [11]. Sau đó, từ thông tin này, một số tham số âm liên quan tới F0 được tính. Các tham số này bao 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. Những tham số này được chọn để phân tích vì sự biến đổi giá trị của chúng thể hiện khá tốt sự thay đổi trong âm điệu và đường cao độ của phát âm. Tiếp theo, hình bao năng lượng được tính theo cách tương tự như cách xác định đường F0. Thông tin năng lượng trước tiên được trích ra bằng cách sử dụng STRAIGHT [77], và sau đó các tham số âm liên quan tới hình bao năng lượng được tính. Các tham số được khảo sát gồm: năng lượng lớn nhất (HPW), năng lượng trung bình (APW), khoảng năng lượng (PWR), năng lượng trung bình của các âm tiết. Sự biến đổi giá trị của những tham số này có thể thể hiện tốt sự thay đổi trong hình bao năng lượng của tín hiệu tiếng nói. 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 vị (bao gồm cả nguyên âm và phụ âm), cũng như khoảng thời gian của khoảng dừng được xác định bằng tay với sự hỗ trợ một phần của Wavesurfer [134]. Hình 4.1 minh họa một ví dụ về việc xác định phân đoạn thời gian.

Trong bảng ở Hình 4.1, dòng đầu tiên chỉ ra các âm vị, trong đó các ô trống tương ứng với khoảng dừng của phát âm. Dòng thứ hai thể hiện thứ tự

âm vị		m	ê		t	í	n		c	ổ	h	ủ
STT âm vị	-1	1	2	0	3	4	5	0	6	7	8	9
thời gian	182	223	331	392	403	476	537	594	616	716	778	895
nguyên âm	-1	2	1	0	2	1	2	0	2	1	2	1

Hình 4.1: Ví dụ về phân đoạn thời gian.

của các âm vị, đánh dấu bằng giá trị -1 trước âm vị đầu tiên; trong dòng này, các khoảng dừng của phát âm được đánh dấu bằng các ô có giá trị bằng 0. Dòng thứ ba chỉ ra thời gian bắt đầu của âm vị tiếp theo cũng như thời gian kết thúc của âm vị hiện tại. Dòng thứ tư thể hiện các âm vị là nguyên âm hay phụ âm: giá trị 1 tương ứng với nguyên âm, giá trị 2 tương ứng với phụ âm, các ô có giá trị bằng 0 tương ứng với các khoảng dừng trong phát âm. Dựa trên bảng phân đoạn thời gian này, 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). Bốn tham số này được lựa chọn vì sự biến đổi của chúng có thể thể hiện hầu hết những thay đổi trong nhịp điệu của phát âm. Cuối cùng, với phổ tín hiệu tiếng nói, các tần số cộng hưởng (F1, F2, F3) và độ nghiêng phổ (ST) được tính. Việc đo tần số cộng hưởng được thực hiện tại điểm giữa của nguyên âm. Tần số lấy mẫu của tín hiệu tiếng nói được giảm về giá trị 10kHz. Phổ thu được bằng cách sử dụng STRAIGHT và ba tần số cộng hưởng F1, F2, F3 được tính với LPC-order 12. Tần số lấy mẫu tín hiệu được giảm xuống 10kHz là vì với tần số này, 5 đỉnh phổ lớn nhất mà LPC cố gắng phân phối theo trục tần số tùy theo sự hiện diện của năng lượng trong dải tần số đặc biệt sẽ được định vị trong khoảng từ 0 đến 5kHz, đây là khoảng tần số quan trọng cho âm thanh tiếng nói, đặc biệt là đối với các nguyên âm. Vị trí của những đỉnh này chính là ước lượng của tần số cộng hưởng. 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.

Sau khi thực hiện giai đoạn trích đặc trưng âm, 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 (trạng thái không cảm xúc) được xác định. Kết quả là chúng ta có 152 tập, mỗi tập chứa 14 giá trị của hệ

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

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

Bảng 4.3: Biến đổi trung bình của các tham số âm của bốn trạng thái cảm xúc so với trạng thái không cảm xúc 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%

số biến đổi. Trong đó có 19 tập cho mỗi một trong bốn trạng thái cảm xúc (vui, buồn, hơi giận, rất giận), 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 nhóm 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. Thực tế, xuất hiện sự khác nhau trong kết quả phân tích giữa hai giọng của cơ sở dữ liệu. Sự khác nhau này là do hai nghệ sĩ thể hiện cảm xúc theo cách khác nhau và với cường độ khác nhau.

Ở mức âm tiết, khi xem xét sự biến đổi của F0, chúng tôi nhận thấy rằng trong cả bốn trạng thái cảm xúc, sự biến đổi của trung bình F0 của các âm tiết tăng dần theo hướng từ đầu tới cuối phát âm, cho cả giọng nam và giọng nữ. Cụ thể hơn, trong các trạng thái cảm xúc vui, hơi giận, và rất giận, càng về cuối phát âm, trung bình F0 của các âm tiết càng tăng so với trạng thái không cảm xúc. Trong khi đó, trong trạng thái cảm xúc buồn, càng về cuối phát âm trung bình F0 của các âm tiết càng giảm so với trạng thái không cảm xúc. Đặc biệt, trung bình F0 của các âm tiết thuộc các từ/cụm từ ở vị trí đầu hoặc cuối câu biến đổi hơn rất nhiều so với trung bình F0 của các âm tiết khác trong câu.

Tương tự như F0, trong cả bốn trạng thái cảm xúc, chúng tôi nhận thấy năng lượng trung bình của các âm tiết cũng có sự biến đổi tăng dần theo hướng từ đầu tới cuối phát âm, cho cả hai giọng trong cơ sở dữ liệu. Chúng ngày càng tăng dần trong các trạng thái cảm xúc vui, hơi giận, rất giận và ngày càng giảm dần trong trạng thái cảm xúc buồn. Đặc biệt, trong trạng thái cảm xúc rất giận, năng lượng trung bình của các âm tiết ở cuối phát âm tăng nhiều hơn đáng kể so với độ tăng của trung bình năng lượng của các âm tiết khác trong phát âm đó. Trong trạng thái cảm xúc vui và trạng thái hơi giận, năng lượng trung bình của các âm tiết thuộc các từ/cụm từ ở vị trí đầu hoặc cuối câu cũng tăng nhiều hơn so với năng lượng trung bình của các âm tiết khác trong câu. Với khoảng thời gian, trong trạng thái cảm xúc vui, với cả hai giọng của cơ sở dữ liệu, khoảng thời gian của hầu hết các âm tiết đều giảm trừ những âm tiết thuộc về từ/cụm từ cuối của phát âm. Những âm tiết này thường có khoảng thời gian dài hơn so với các âm tiết tương ứng trong trạng thái không cảm xúc, đặc biệt khi chúng kết thúc bởi nguyên âm/bán nguyên âm hay khi chúng là âm tiết có thanh điệu. Sự kéo dài của các âm tiết cuối trong trạng thái cảm xúc vui xảy ra chủ yếu ở phần cuối của âm tiết. Trong khi đó, với trạng thái cảm xúc rất giận, khoảng thời gian của các âm tiết đều giảm, đặc biệt là đối với các âm tiết thuộc từ/cụm từ cuối phát âm; các âm tiết thuộc từ/cụm từ cuối phát âm có khoảng thời gian giảm hơn rất nhiều so với trạng thái không cảm xúc. 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. AP là viết tắt của trung bình F0, APW là viết tắt của năng lượng trung bình, MD là viết tắt của khoảng thời gian trung bình, tiền tố F chỉ các âm tiết đầu phát âm, L chỉ âm tiết cuối phát âm.

4.4 Tổng hợp tiếng nói tiếng Việt có cảm xúc

4.4.1 Xây dựng luật biến đổi tiếng nói tiếng Việt không cảm xúc thành tiếng nói có cảm xúc

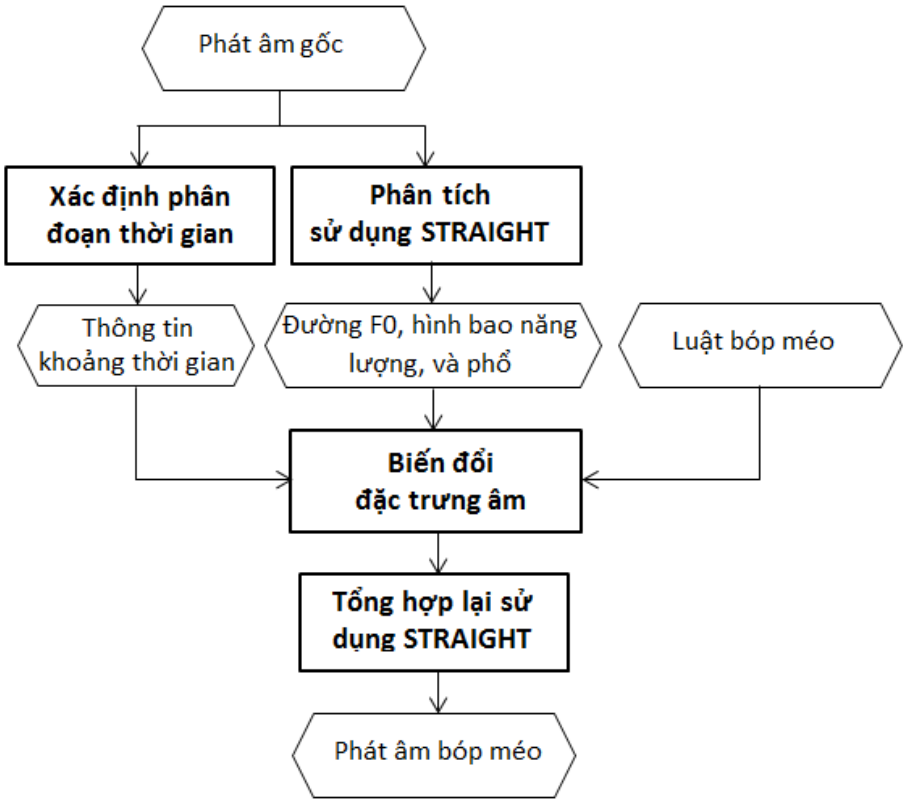
Có một thực tế rằng khi trạng thái cảm xúc thay đổi, đặc trưng âm biến đổi nhiều hơn ở một số âm tiết thay vì biến đổi đồng đều trong tất cả các âm

tiết. Trong các ngôn ngữ đa âm tiết (ví dụ tiếng Anh, tiếng Nhật), những âm tiết này thường là các âm tiết trọng âm. Tuy nhiên, trong hầu hết các nghiên cứu về tổng hợp tiếng nói có cảm xúc đã được công bố, đặc trưng âm được phân tích và biến đổi ở mức phát âm của câu, điều này có thể sẽ làm giảm tính tự nhiên của tiếng nói tổng hợp có cảm xúc. Tiếng Việt là ngôn ngữ đơn âm tiết và không có trọng âm. Mặc dù vậy, 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 không cảm xúc 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 không cảm xúc 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 (Các luật cho các cảm xúc khác được xây dựng theo cách tương tự):

(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)

Có thể diễn giải luật (1) như sau: Để biến đổi một phát âm tiếng Việt ở trạng thái không cảm xúc thành phát âm có cảm xúc vui thì tần số cao nhất (HP) tăng 12.23%, tần số (AP) của các âm tiết thường (không phải là âm tiết đầu/cuối của phát âm) tăng 7.75%, khoảng tần số (PR) tăng 51.57%, năng lượng (APW) của các âm tiết thường tăng 17.21%, năng lượng cao nhất (HPW) tăng 7.96%, khoảng năng lượng (PWR) tăng 12.61%, khoảng dừng (MPAU) giảm 3%, khoảng thời gian của các phụ âm (CL) giảm 3.15%, tỉ lệ giữa thời gian của phụ âm và nguyên âm (RCV) giảm 10.24%, tổng thời gian (TL)giảm 3.55%, tần số F1 tăng 9.99%, F2 tăng 15.43%, F3 tăng 2.17%, chỉ số ST giảm 14%, tần số của âm tiết thuộc từ/cụm từ đầu phát âm (F-AP) tăng 8.35%, năng lượng của âm tiết thuộc từ/cụm từ đầu phát âm (F-APW) tăng 17.42%, khoảng thời gian của âm tiết thuộc từ/cụm từ đầu phát âm (F-MD) tăng 1.05%, tần số của âm tiết thuộc từ/cụm từ cuối phát âm (L-AP) tăng 9.05%, năng lượng của âm

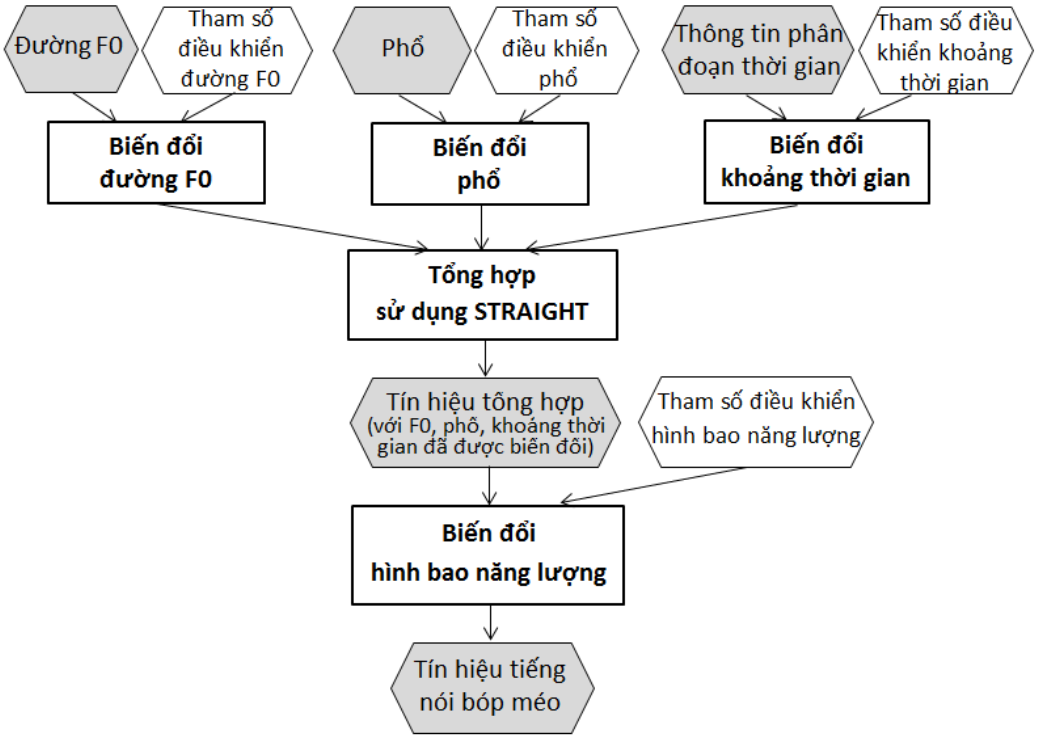


Hình 4.2: Tiến trình biến đổi tiếng nói sử dụng STRAIGHT

tiết thuộc từ/cụm từ cuối phát âm (L-APW) tăng 19.23%, khoảng thời gian của âm tiết thuộc từ/cụm từ cuối phát âm (L-MD) tăng 19.84%. Với luật này, đặc trưng âm được biến đổi không đồng đều ở các âm tiết. Ví dụ, để tạo phát âm ở trạng thái cảm xúc vui, tần số (AP) của các âm tiết đều tăng, nhưng tần số của các âm tiết ở đầu/cuối phát âm được điều chỉnh tăng nhiều hơn so với các âm tiết khác; ví dụ khác là 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

Để tổng hợp tiếng nói có cảm xúc bằng phương pháp điều khiển tường minh thì các tham số âm liên quan đến cảm xúc cần phải được kết hợp với tiếng nói ở trạng thái không cảm xúc theo tập các luật thu được từ giai đoạn phân



Hình 4.3: Tiến trình biến đổi đặc trưng âm.

tích cơ sở dữ liệu. Quá trình kết hợp này được thực hiện trên yếu tố ngôn điệu cũng như âm sắc. Trong nghiên cứu của luận án, kỹ thuật biến đổi tiếng nói được sử dụng để tạo ra tiếng nói tiếng Việt có cảm xúc. Tiến trình biến đổi tiếng nói được thể hiện trong Hình 4.2.

Trước tiên, STRAIGHT [77] được dùng để trích ra đường F0, hình bao năng lượng, và phổ của tín hiệu tiếng nói không cảm xúc, 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 suy ra từ tập các hệ số biến đổi 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 Phần 4.3.2 và 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.

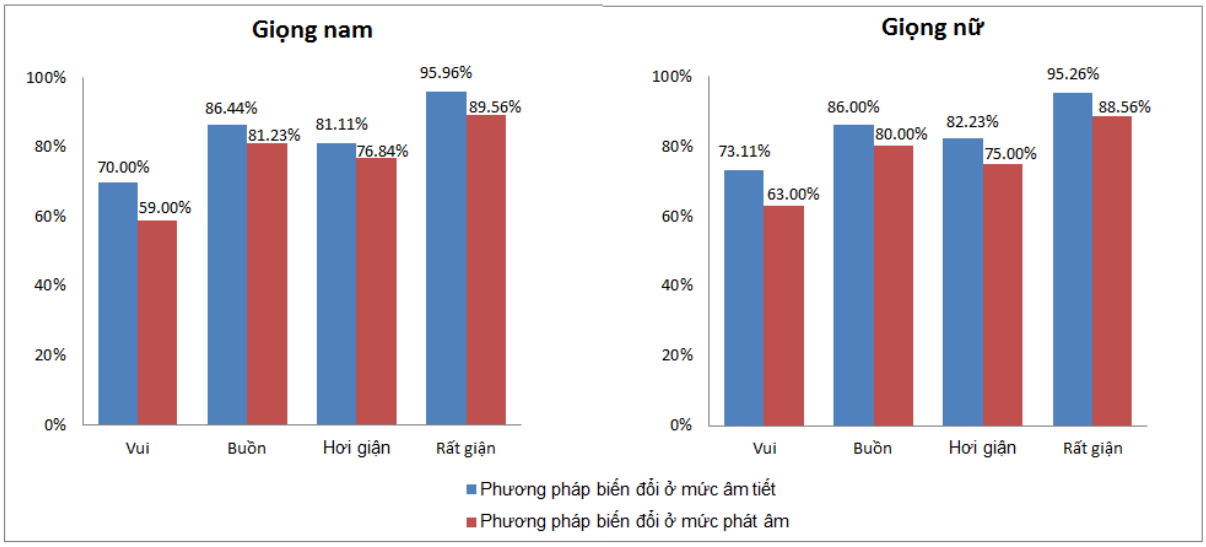
4.5 Thực nghiệm và đánh giá

Sử dụng phương pháp được trình bày trong Phần 4.4, chúng tôi đã tiến hành thực nghiệm để tổng hợp tiếng nói tiếng Việt có cảm xúc từ tiếng nói không cảm xúc. Trước tiên, 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; 10 câu này cũng có đặc điểm là hầu như không chứa nội dung ý nghĩa cảm xúc. Sau đó, các phát âm ở trạng thái không cảm xúc của 10 câu vừa nêu được tạo bởi 1 nam và 1 nữ (không phải là hai nghệ sĩ Việt được nói đến ở Phần 4.3.1). Các phát âm ở trạng thái không cảm xúc 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, để so sánh, đánh giá kết quả của phương pháp biến đổi ở mức âm tiết, các luật được suy ra **chỉ** từ Bảng 4.2 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 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, với cả hai phương pháp tổng hợp tiếng nói, 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 trong Hình 4.4. Thực nghiệm 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.

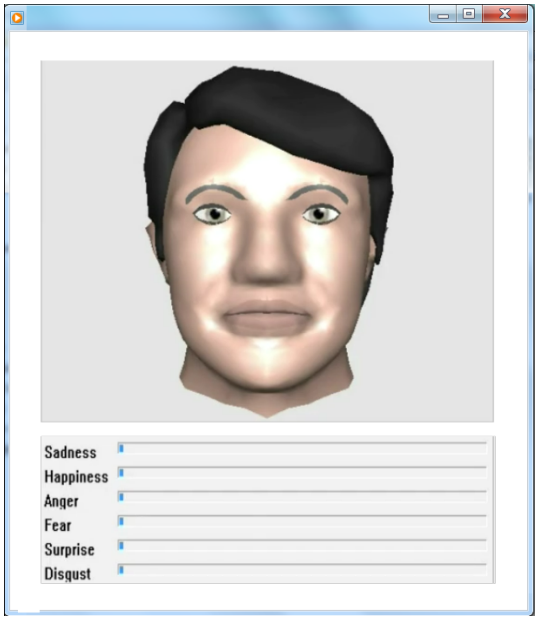
Kết quả này cho thấy cơ chế tổng hợp tiếng nói khá hiệu quả, và các luật được sử dụng khá phù hợp.

Để đánh giá khả năng của mô hình tổng hợp tiếng nói tiếng Việt (được trình bày ở Phần 4.4) trong việc tạo biểu cảm giọng điệu thể hiện cảm xúc khi áp dụng cho nhân vật ảo, chúng tôi đã tiến hành thực nghiệm để thu thập đánh giá của người dùng. Chúng tôi đã áp dụng mô hình tổng hợp này cho nhân vật ảo Obie được đề cập trong Phần 3.4 ở Chương 3 của luận án. Quá trình tiến hành thực nghiệm và kết quả đánh giá như sau:

Đối tượng được đánh giá: Nhằm đánh giá ưu điểm và hiệu quả của mô hình đề xuất, 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, tiếng nói của nhân vật ảo A là tiếng nói ở trạng thái không cảm xúc.
- Nhân vật ảo B: chính là một bản sao của nhân vật ảo A, nhưng ở đây bản sao này đã được áp dụng "Phương pháp biến đổi ở mức phát âm" 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 bản sao này đã được áp dụng "Phương pháp biến đổi ở mức âm tiết" để tạo biểu cảm giọng điệu cho nhân vật ảo C.

Chuẩn bị cho thực nghiệm đánh giá:



Hình 4.5: Hình ảnh minh họa video clip dùng để đánh giá mô hình tạo biểu cảm giọng điệu.

Để tiến hành thực nghiệm đánh giá, chúng tôi xây dựng cho mỗi nhân vật ảo 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. Hình ảnh của video clip được minh họa trong Hình 4.5.

Mục tiêu của thực nghiệm đó là đánh giá tính thuyết phục của các nhân vật ảo A, B, C trong việc tạo biểu cảm giọng điệu thể hiện trạng thái cảm xúc được chỉ ra ở phần dưới trong hình ảnh của video clip. Những người tham gia thực nghiệm sẽ đánh giá xem các nhân vật ảo có thể hiện trong giọng nói đúng trạng thái cảm xúc được chỉ ra hay không, cách thể hiện cảm xúc có tự nhiên và hợp lý không.

Tiếp đến, phương pháp ghi lại kết quả đánh giá của người dùng được xây dựng. 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, theo mẫu được chỉ ra trên Hình 4.6.

Kịch bản tiến hành thực nghiệm:

Thực nghiệm đánh giá tính thuyết phục của các nhân vật ảo A, B, C trong việc tạo biểu cảm giọng điệu thể hiện cảm xúc được tiến hành với 14 người tham

Tính thuyết phục của nhân vật ảo thứ nhất?						
0 (Rất kém)	**1** (Kém)	**2** (Hơi kém)	**3** (Trung bình)	**4** (Khá tốt)	**5** (Tốt)	**6** (Rất tốt)
Tính thuyết phục của nhân vật ảo thứ hai?						
0 (Rất kém)	**1** (Kém)	**2** (Hơi kém)	**3** (Trung bình)	**4** (Khá tốt)	**5** (Tốt)	**6** (Rất tốt)
Tính thuyết phục của nhân vật ảo thứ ba?						
0 (Rất kém)	**1** (Kém)	**2** (Hơi kém)	**3** (Trung bình)	**4** (Khá tốt)	**5** (Tốt)	**6** (Rất tốt)

Hình 4.6: Mẫu ghi kết quả đánh giá tính thuyết phục trong việc thể hiện cảm xúc trong giọng nói của các nhân vật ảo .

gia trong độ tuổi từ 20 đến 35, có khả năng nghe, nói, nhìn bình thường. Thực nghiệm được tiến hành trong phòng kín, cách âm tương đối tốt nhằm hạn chế tối đa ảnh hưởng của các yếu tố bên ngoài đối với kết quả đánh giá của người dùng. Mỗi phiên thực nghiệm được tiến hành riêng cho từng người như sau: Trước tiên người dùng được giới thiệu về video clip, về mục tiêu của thực nghiệm, về mẫu ghi kết quả đánh giá. Tiếp đến, người dùng sẽ lần lượt xem ba video clip về ba nhân vật ảo đã đề cập ở trên (thứ tự ngẫu nhiên); số lần xem video clip không bị giới hạn, người dùng có thể yêu cầu xem lại nếu họ muốn. Sau khi đã xem video clip, người dùng được yêu cầu ghi kết quả đánh giá vào mẫu như Hình 4.6. Với mỗi nhân vật ảo, người dùng sẽ khoanh tròn vào số điểm họ chọn.

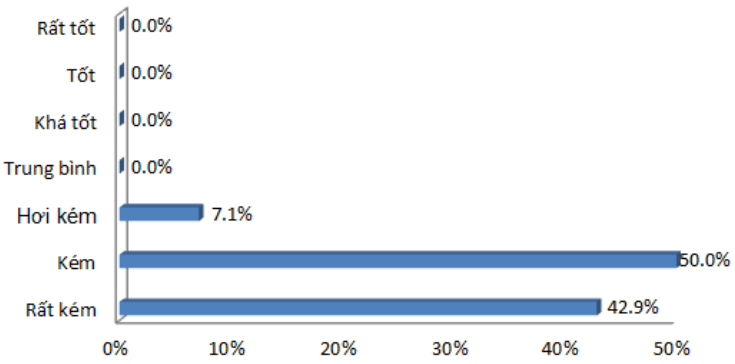
Kết quả đánh giá:

Sau khi tiến hành thực nghiệm, kết quả đánh giá của người dùng 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 (điều này là hiển nhiên bởi vì thực tế tiếng nói của nhân vật ảo A hoàn toàn ở trạng thái không cảm xúc), 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.

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.

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

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



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

Xét cặp giả thuyết, đối thuyết:

$$H_0 : \mu_B - \mu_C \geq 0,$$

$$H_1 : \mu_B - \mu_C < 0$$

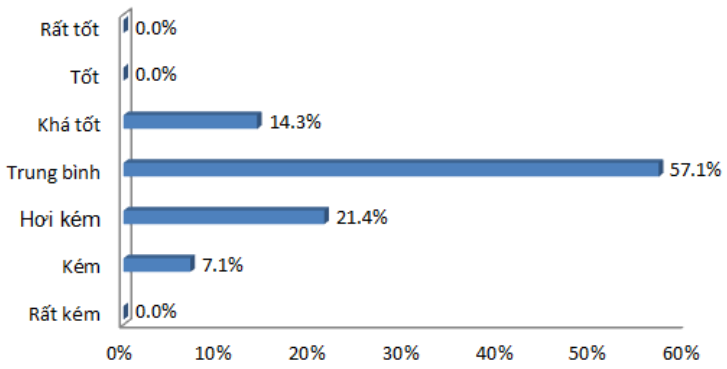
Chúng tôi chọn mức ý nghĩa là 0.05 và sử dụng phương pháp kiểm định *matched-pairs t-test*.

Đặt $D = B - C$, $D_i = B_i - C_i$, D_i nhận các giá trị sau:

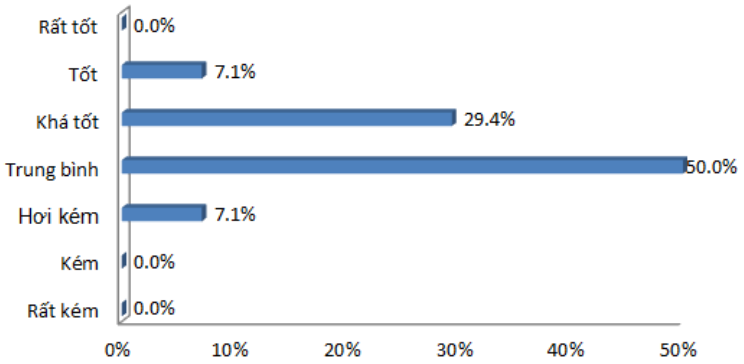
$$D_i : -1 \quad -1 \quad 1 \quad 0 \quad -1 \quad -1 \quad 0 \quad 0 \quad 0 \quad -2 \quad -1 \quad 0 \quad -1 \quad -2$$

Ta có $\bar{d} = -0.64286$.

Từ đó, độ lệch chuẩn của sự khác nhau được tính từ các cặp đôi là:



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



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

$$s_d = sqrt[(\Sigma(d_i - \overline{d})^2/(n-1)] = 0.84190.$$

Lỗi chuẩn của phân phối lấy mẫu của $\overline{d}$ là

$$SE = s_d/sqrt(n) = 0.22501.$$

Độ tự do $DF = n - 1 = 13$.

Từ đó tính được $t = [(\overline{x_1} - \overline{x_2}) - D]/SE = (\overline{d} - D)/SE = -2.85706$.

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

Vì $P = 0.00674 < 0.05$ nên giả thuyết H_0 bị từ chối; trung bình điểm đánh giá tính thuyết phục của nhân vật ảo C (3.429) lớn hơn về mặt thống kê so với trung bình điểm đánh giá tính thuyết phục của nhân vật ảo B (2.786). Từ kết quả này, 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* 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 thể hiện cảm xúc cho nhân vật ảo nói tiếng Việt.

4.6 Kết chương

Chương 4 của luận án đã đề xuất mô hình biến đổi tiếng nói tiếng Việt từ trạng thái không cảm xúc thành tiếng nói có cảm xúc, cung cấp cho nhân vật ảo khả năng thể hiện cảm xúc trong giọng nói tiếng Việt. Tiếng nói ở trạng thái không cảm xúc được biến đổi bằng cách sử dụng các luật biến đổi liên quan đến tần số cơ bản, thời gian, năng lượng, phổ. Các luật này được xây dựng từ kết quả phân tích cơ sở dữ liệu tiếng nói tiếng Việt có cảm xúc. Tiếng nói không cảm xúc được biến đổi có tính đến sự biến đổi của đặc trưng âm ở mức âm tiết; đặc trưng âm được biến đổi nhiều hơn ở một số âm tiết thay vì biến đổi đồng đều trong tất cả các âm tiết của câu. Đây là điểm khác so với các nghiên đã được đề xuất, khiến cho tiếng nói tổng hợp có cảm xúc tự nhiên hơn, thuyết phục hơn. Kết quả của thực nghiệm đánh giá đã khẳng định điều này; các trạng thái cảm xúc tổng hợp được nhận dạng tương đối tốt. Kết quả này chỉ ra rằng mô hình biến đổi tiếng nói được đề xuất có thể được sử dụng cho nhân vật ảo nói tiếng Việt nhằm tăng cường khả năng thể hiện cảm xúc của chúng.

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

Các nhà tâm lý học cũng như các nhà nghiên cứu trong các lĩnh vực khác từ lâu đã nhận ra tầm quan trọng của biểu cảm khuôn mặt trong việc thể hiện và đánh giá cảm xúc; kênh thông tin này có thể nhận được sự quan tâm bằng tất cả các kênh khác cộng lại. Kênh thông tin quan trọng thứ hai trong việc thể hiện và đánh giá cảm xúc là tiếng nói; "rất nhiều biến đổi trong hành xử phát âm có thể tạo nên sự cảm nhận khác nhau" [24]. Vì vậy, luận án tập trung vào hai kênh này khi giải quyết bài toán thể hiện cảm xúc cho nhân vật ảo nói tiếng Việt.

Mô hình tạo biểu cảm khuôn mặt thể hiện trạng thái cảm xúc liên tục đã được trình bày trong Chương 3, và mô hình tạo biểu cảm giọng điệu thể hiện cảm xúc trong giọng nói tiếng Việt đã được mô tả ở Chương 4. Trong chương này của luận án, chúng tôi 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. Khuôn mặt ba chiều được xây dựng dựa trên các kết quả được trình bày trong Chương 3, Chương 4 và áp dụng một số nghiên cứu đã được công bố bởi các tác giả khác trong cùng lĩnh vực. Khuôn mặt ba chiều này có khả năng thể hiện cảm xúc trên khuôn mặt và trong giọng nói tiếng Việt một cách tự nhiên. Nó có khả năng thể hiện cử động của môi trong khi phát âm các từ tiếng Việt, và cùng lúc đó cũng có khả năng thể hiện biểu cảm khuôn mặt trong khi nói. Theo hiểu biết của chúng tôi, cho tới nay chưa có khuôn mặt nào như vậy được xây dựng.

Hệ thống khuôn mặt gồm ba mô đun: mô đun tạo biểu cảm khuôn mặt thể

hiện cảm xúc liên tục - sử dụng kết quả nghiên cứu trong Chương 3; mô đun tạo biểu cảm giọng điệu trong giọng nói tiếng Việt - sử dụng kết quả nghiên cứu trong Chương 4; và một mô đun nữa có chức năng tạo chuyển động của môi khi phát âm các từ tiếng Việt, và kết hợp các chuyển động này với cử động khuôn mặt thể hiện cảm xúc. Để kiểm tra, đánh giá khuôn mặt được xây dựng, chúng tôi đặt nó trong miền của cổ động viên bóng đá; khuôn mặt được sử dụng làm khuôn mặt của một cổ động viên bóng đá ảo, cổ động viên này trải qua các cảm xúc khác nhau và thể hiện cảm xúc đó trên khuôn mặt cũng như trong giọng nói của anh ta.

Nội dung của chương được tổ chức như sau. Phần 5.2 trình bày nghiên cứu của các tác giả khác, được áp dụng để xây dựng khuôn mặt ba chiều nói tiếng Việt. Sau đó, Phần 5.3 trình bày về kiến trúc của hệ thống khuôn mặt ba chiều nói tiếng Việt. Trong phần này, quá trình xây dựng và hoạt động của ba mô đun chính của hệ thống sẽ được mô tả chi tiết. Tiếp theo, thực nghiệm và đánh giá được trình bày ở Phần 5.4. Cuối cùng là phần kết luận chương.

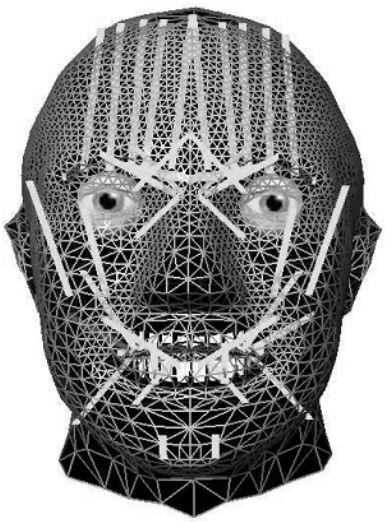
5.2 Những nghiên cứu liên quan

Phần này đề cập đến nghiên cứu của các tác giả khác, được luận án áp dụng để xây dựng khuôn mặt ba chiều nói tiếng Việt.

Mô hình khuôn mặt ba chiều

Để xây dựng khuôn mặt ba chiều nói tiếng Việt, luận án áp dụng mô hình khuôn mặt ba chiều được đề xuất bởi tác giả Bui và cộng sự [15]. Đây là mô hình khuôn mặt ba chiều dựa trên cơ (muscle-based) có khả năng tạo cử động khuôn mặt tự nhiên với chất lượng cao trong thời gian thực trên máy tính cá nhân bình thường. Mô hình khuôn mặt được minh họa trên Hình 5.1, 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.

Dữ liệu lưới khuôn mặt ban đầu có được từ việc sử dụng máy quét 3D, sau đó được xử lý qua hai giai đoạn: giai đoạn 1 giảm số đỉnh và số đa giác tại những vùng ít biểu cảm trên khuôn mặt, khiến tốc độ hoạt họa tăng mà vẫn đảm bảo độ mịn, chi tiết tại những vùng biểu cảm trên khuôn mặt (sau giai đoạn 1 lưới



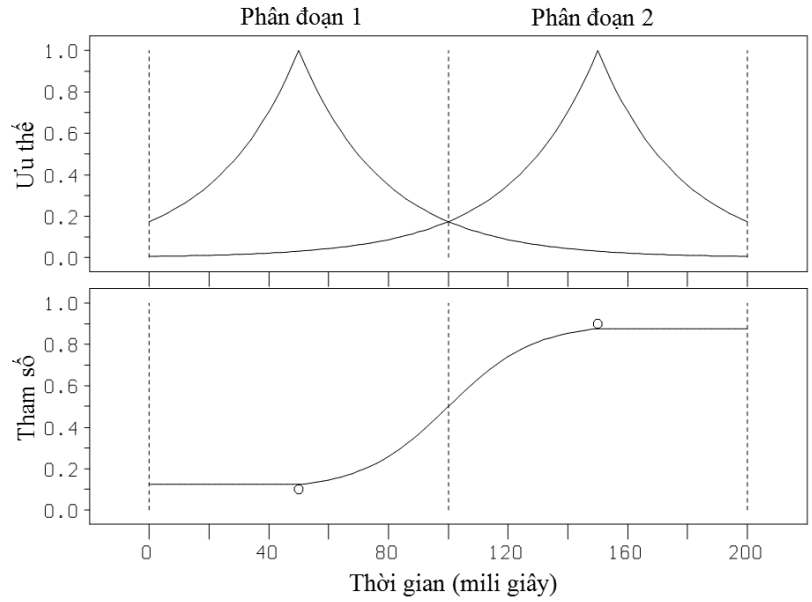
Hình 5.1: Mô hình khuôn mặt 3D đề xuất bởi Bui và cộng sự [15].

gồm 2480 đỉnh và 4744 đa giác); ở giai đoạn 2, dựa trên phân bố cơ mặt người chia khuôn mặt thành 11 vùng nhằm giới hạn và điều khiển sự di chuyển của các đỉnh đa giác do sự co của các cơ. Mô hình môi sử dụng một mặt B-spline với lưới 24x6 điểm điều khiển để đảm bảo sự mịn khi thay đổi hình dáng môi do hoạt động của cơ.

Sau khi khuôn mặt đã được mô hình hóa, cử động trên khuôn mặt được tạo ra bởi sự điều khiển của các cơ. Bui và cộng sự cài đặt các cơ giả để mô phỏng cơ thực điều khiển hoạt động của miệng và mắt; các vùng khác của khuôn mặt được điều khiển bởi cơ véc tơ; ngoài ra, các tác giả còn mô phỏng sự quay của quai hàm. Mỗi cơ có một vùng ảnh hưởng, và khi co sẽ khiến các đỉnh đa giác nằm trong vùng ảnh hưởng đó dịch chuyển vị trí, từ đó tạo ra cử động trên khuôn mặt. Các tác giả đã đưa ra giải pháp để tổng hợp sự co của nhiều cơ bằng cách mô phỏng trạng thái song song của chúng. Với một đỉnh đa giác nằm trong vùng ảnh hưởng của nhiều cơ, mức co của các cơ được chia nhỏ ra để áp dụng cho đỉnh đó; với mỗi lần áp dụng, sự dịch chuyển vị trí của đỉnh đa giác gây ra bởi sự co ở mức nhỏ của các cơ được cộng lại.

Tạo chuyển động của môi khi phát âm tiếng nói

Theo Cohen và Massaro [23], hình dáng môi khi phát âm tiếng nói có tầm quan trọng ngang với thông tin âm thanh của tiếng nói. Vì vậy, tạo chuyển động của môi khi phát âm là thao tác cần thiết nhằm tăng cường tính thuyết phục



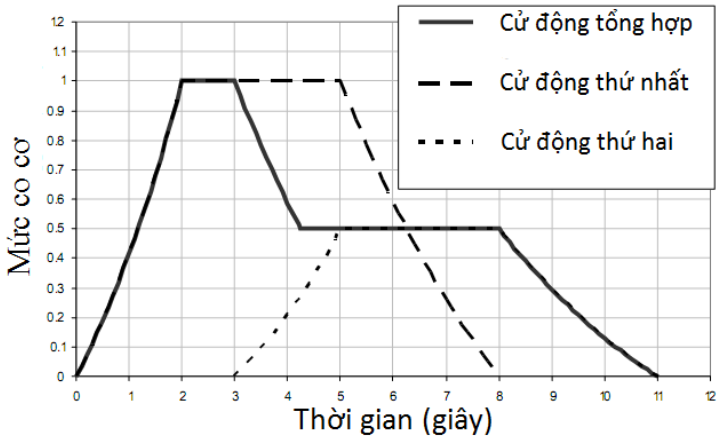
Hình 5.2: Ưu thế của hai phân đoạn tiếng nói theo thời gian (hình trên) và hàm tham số điều khiển sau khi áp dụng hiệu ứng đồng phát âm đề xuất bởi Cohen và Massaro [23] (hình dưới)

của nhân vật ảo. Trong nghiên cứu [23], các tác giả đã 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.

Trong [23], 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 (dominance) khi phát âm, ưu thế này tăng và giảm dần theo thời gian trong quá trình phát âm. Hàm ưu thế (dominance function) 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. Các hàm ưu thế có thể chồng nhau trong một khoảng thời gian cho trước; trung bình có trọng số của tất cả các hàm ưu thế sẽ tạo ra hình dáng cuối cùng của môi. Hình 5.2 minh họa ví dụ hàm ưu thế và hoạt động của một tham số sau khi áp dụng hiệu ứng đồng phát âm.

Tổng hợp các cử động trên khuôn mặt

Có nhiều loại cử động khuôn mặt khác nhau, như tín hiệu giao tiếp, thể hiện cảm xúc, cử động của môi khi nói,... Các loại cử động khác nhau trên khuôn



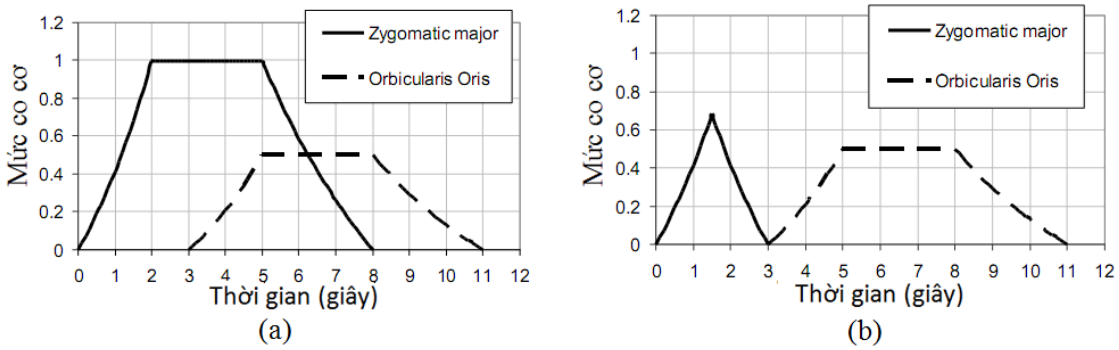
Hình 5.3: Cơ chế tổng hợp cử động trong cùng một kênh [17]

mặt có thể xảy ra đồng thời, ví dụ như vừa phát âm tiếng nói vừa thể hiện cảm xúc; vấn đề đặt ra là làm thể nào để tổng hợp các cử động xảy ra đồng thời này để tạo hoạt họa khuôn mặt tự nhiên cho nhân vật ảo.

Trong nghiên cứu [17], các tác giả đã đề xuất cơ chế tổng hợp các loại cử động khuôn mặt khác nhau trên mô hình khuôn mặt 3D được đề xuất trong [15]. Cơ chế này có khả năng tích hợp các loại cử động để tạo hoạt họa tự nhiên trên khuôn mặt. Cử động khuôn mặt được chia thành các nhóm gọi là các kênh, tùy theo loại cử động, ví dụ như cử động thể hiện cảm xúc, cử động của môi khi nói... Sau đó 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. Cơ chế này được minh họa trên Hình 5.3, thể hiện quá trình tổng hợp hai cử động liên quan đến cơ Zymgomatic Major; cử động tổng hợp đi theo cử động thứ nhất cho tới giây thứ 3, đây là thời điểm xuất hiện cử động thứ hai, sau đó cử động tổng hợp giảm dần về giá trị đích của cơ Zymgomatic Major ở cử động thứ hai (0.5), và tiếp theo thì đi theo cử động thứ hai.

Để tổng hợp cử động từ các kênh khác nhau, trước tiên 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

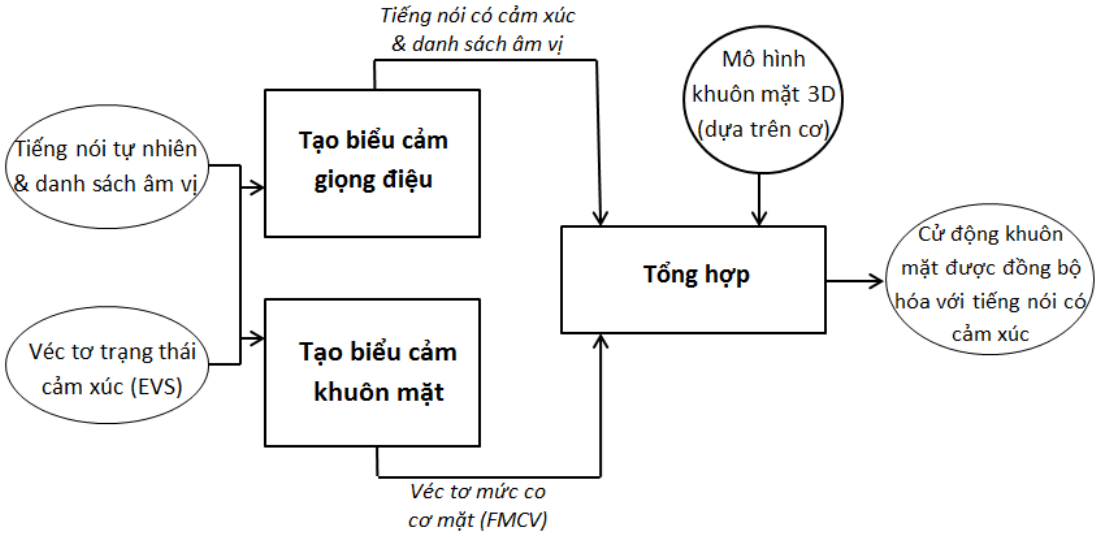


Hình 5.4: Cơ chế tổng hợp cử động hai kênh khác nhau [17]. (a): Hai cử động trước khi tổng hợp; (b): Cử động sau khi áp dụng cơ chế tổng hợp.

giá trị lớn nhất của tham số đó từ tất cả các kênh. Tại thời điểm t, khi có xung đột xảy ra giữa các tham số ở các kênh khác nhau thì tham số liên quan tới cử động với độ ưu tiên cao hơn sẽ chiếm ưu thế và lấn át tham số với độ ưu tiên thấp hơn. Hoạt động của tham số bị lấn át tại khoảng thời gian xung quanh t sẽ được điều chỉnh để nó không tăng/giảm quá nhanh. Hình 5.4 minh họa cơ chế tổng hợp cử động từ các kênh khác nhau. Cơ Orbicularis Oris liên quan tới cử động của môi khi phát âm, cơ này xung đột với cơ Zygomatic Major và có độ ưu tiên cao hơn. Khi cơ Orbicularis Oris được kích hoạt (tại giây thứ 3) thì cơ Zygomatic Major bị chặn; hoạt động của cơ Zygomatic Major tại khoảng thời gian trước giây thứ 3 sẽ được điều chỉnh để nó không giảm quá nhanh.

5.3 Kiến trúc hệ thống

Khuôn mặt ba chiều nói tiếng Việt được xây dựng chủ yếu dựa trên các kết quả nghiên cứu ở Chương 3 và Chương 4. Kiến trúc tổng thể của hệ thống được minh họa trên Hình 5.5. Đầu vào của hệ thống là chuỗi các véc tơ trạng thái cảm xúc theo thời gian (EVS) và tiếng nói ở trạng thái không cảm xúc cùng với danh sách các âm vị tương ứng có kèm theo thông tin thời gian. Theo cách hoàn hảo hơn thì một phần của đầu vào nên là text thay vì tiếng nói không cảm xúc kèm theo danh sách âm vị. Nhưng nội dụng nghiên cứu của luận án chỉ tập trung vào bài toán biến đổi tiếng nói tiếng Việt để thể hiện cảm xúc; luận án không giải quyết bài toán tổng hợp tiếng nói từ text. Vì vậy, chúng tôi giả sử rằng đã có sẵn một hệ thống tổng hợp tiếng nói tiếng Việt từ text, và đầu ra của hệ thống này được sử dụng làm đầu vào cho hệ thống (khuôn mặt) được



Hình 5.5: Kiến trúc hệ thống khuôn mặt 3D nói tiếng Việt.

xây dựng.

Hệ thống khuôn mặt có ba mô đun chính: mô đun *Tạo biểu cảm giọng điệu (VESS)*, mô đun *Tạo biểu cảm khuôn mặt(EFE)*, và mô đun *Tổng hợp*. Mô đun *VESS* sử dụng hệ thống trong Chương 4 để chuyển tiếng nói tiếng Việt ở trạng thái không cảm xúc thành tiếng nói có cảm xúc. Mô đun *EFE* sử dụng hệ thống trong Chương 3 để mô phỏng biểu cảm khuôn mặt thể hiện cảm xúc liên tục từ chuỗi các véc tơ trạng thái cảm xúc (ESV). Từ danh sách các âm vị kèm theo thông tin thời gian, 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. Cuối cùng, biểu cảm khuôn mặt và các chuyển động sẽ được hiển thị đồng bộ hóa với tiếng nói có cảm xúc trên một khuôn mặt ba chiều. Luận án sử dụng mô hình khuôn mặt ba chiều dựa trên cơ, được đề xuất trong nghiên cứu [15]. Mô hình khuôn mặt này có khả năng tạo biểu cảm khá tự nhiên và trong thời gian thực trên các máy tính cá nhân thông thường. Quá trình xây dựng và hoạt động của các mô đun được trình bày trong ba phần nhỏ tiếp theo.

5.3.1 Mô đun *Tạo biểu cảm giọng điệu* (Vietnamese Emotional Speech Synthesis - 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 không cảm xúc 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. Từ đầu vào là tiếng nói không cảm xúc kèm theo danh sách âm vị và thông tin thời gian, mô đun VESS sẽ tạo ra tiếng nói tiếng Việt có cảm xúc tương ứng; quá trình này được thực hiện theo tiến trình minh họa trong Hình 4.2 ở Chương 4. Cụ thể, từ phát âm đầu vào, STRAIGHT [77] được dùng để trích ra đường F0, hình bao năng lượng, và phổ của tín hiệu tiếng nói không cảm xúc. Sau đó đặc trưng âm liên quan tới F0, năng lượng, phổ, và thời gian được biến đổi dựa trên các luật suy ra từ quá trình phân tích cơ sở dữ liệu được trình bày trong Phần 4.3.2 ở Chương 4. 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. Tiếng nói có cảm xúc này cùng với danh sách âm vị và thông tin thời gian sẽ là một phần đầu vào của mô đun *Tổng hợp*.

5.3.2 Mô đun *Tạo biểu cảm khuôn mặt* (Emotions to Facial Expressions - 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 mức co cơ mặt (FMCV) tương ứng. Mô đun này sử dụng mô hình được minh họa trong Hình 3.10 ở Chương 3 để chuyển các trạng thái cảm xúc liên tục của nhân vật ảo thành mức co cơ tạo ra các biểu cảm tương ứng với cảm xúc đầu vào. Ý tưởng chính của mô hình này đó là khi một cảm xúc được kích hoạt, biểu cảm khuôn mặt tương ứng xảy ra theo chuỗi với cường độ giảm dần. Ví dụ, khi có một sự kiện nào đó kích hoạt cảm xúc vui của nhân vật ảo thì nhân vật này sẽ không cười với cường độ lớn trong suốt khoảng thời gian mà cảm xúc vui tồn tại; thay vào đó nó sẽ thể hiện một chuỗi các biểu cảm cười với cường độ giảm dần. Đầu ra của mô đun EFE (chuỗi véc tơ độ co cơ mặt) sẽ

là một phần đầu vào của mô đun *Tổng hợp*.

5.3.3 Mô đun *Tổng hợp*

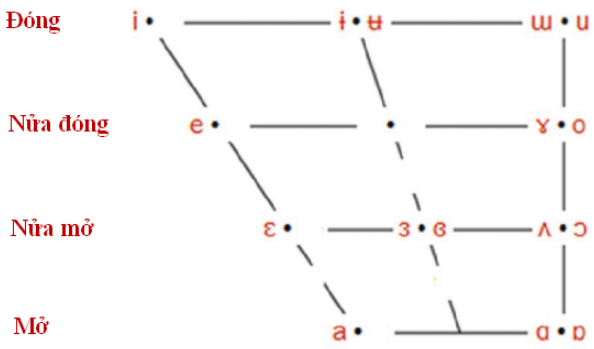
Mô đun *Tổng hợp* tạo chuyển động của môi khi phát âm tiếng Việt và kết hợp các chuyển động này với cử động khuôn mặt thể hiện cảm xúc.

Hình vị cho các âm vị tiếng Việt

Để tạo chuyển động của môi khi phát âm các từ tiếng Việt, trước tiên chúng ta cần có một tập các hình vị cho khuôn mặt, tương ứng với các âm vị tiếng Việt. Luận án dựa trên các luật được đưa ra trong các nghiên cứu [1] và [106] để xác định hình vị tương ứng của mỗi âm vị tiếng Việt.

Theo [1], âm vị tiếng Việt được chia thành hai loại: nguyên âm và phụ âm. Với hình vị của các nguyên âm, những âm vị này được phân chia và thể hiện tùy theo ba yếu tố chính: vị trí của lưỡi, độ mở của miệng, và hình dáng của môi. Với yếu tố độ mở của miệng, các nguyên âm được chia thành bốn loại: nguyên âm đóng (*i*), nguyên âm nửa đóng (*ê*), nguyên âm nửa mở (*e*), và nguyên âm mở(*a*). Tính chất hẹp - rộng của nguyên âm được xác định bởi độ mở dần của miệng. Với hình dáng của môi, các nguyên âm được chia thành hai loại: nguyên âm tròn môi (*o, ô*) và nguyên âm không tròn môi (*ơ*). Tính chất tròn môi hay không tròn môi của nguyên âm được quyết định bởi hình dáng của môi. Hình 5.6 thể hiện mối quan hệ giữa các nguyên âm và hai yếu tố nói trên. Các đường ngang thể hiện độ mở của miệng. Các đường dọc thể hiện hình dáng của môi; phần bên trái chỉ ra các nguyên âm không tròn môi, phần bên phải chỉ ra các nguyên âm tròn môi. Với hình vị của các phụ âm, những âm vị này được phân chia và thể hiện tùy theo hai yếu tố chính: âm vị được phát âm ở đâu và được phát âm như thế nào. Theo yếu tố đầu tiên, các phụ âm được chia thành ba loại: phụ âm môi (*b, p,v, ph*), phụ âm lưỡi (*đ, ch, c,k*), và phụ âm họng (*h*).

Vì mô hình khuôn mặt 3D mà luận án sử dụng [15] mô phỏng hoạt động của cơ véc tơ, cơ điều khiển mắt, miệng, và sự quay của quai hàm nên nó có thể thể hiện chuyển động của môi khi phát âm tiếng Việt. Độ mở của miệng tương ứng với lượng quay của quai hàm; và độ tròn của môi phụ thuộc vào các cơ có ảnh hưởng lên môi. Để đơn giản, một số nguyên âm tương đối giống nhau được đưa vào cùng một nhóm. Để tạo hình vị cho các nguyên âm, lượng quay của



Hình 5.6: Hình thang nguyên âm.

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, chúng tôi chỉ quan tâm tới vị trí mà âm vị được phát âm. Theo yếu tố này, các phụ âm được chia thành ba loại: phụ âm môi - môi, phụ âm môi - răng, và loại thứ ba chứa các phụ âm còn lại. Chúng tôi áp dụng các luật trong [1] và [106] 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 cử động của môi khi phát âm tiếng Việt

Lời nói của con người thường là các đoạn, câu, hoặc một số từ. Những đơn vị này bao gồm một tập các âm vị, một số âm vị kết hợp với nhau sẽ tạo thành một từ. Với mỗi âm vị đơn chúng ta đã có một hình vị tương ứng. Bây giờ yêu cầu tiếp theo là tạo sự chuyển đổi từ một hình vị (ví dụ V1) sang một hình vị khác (ví dụ V2) một cách từ từ và mịn nhằm tạo ra chuyển động tự nhiên của môi khi nói. Cách đơn giản nhất đó là tạo các hình vị trung gian của V1 và V2 bằng cách cộng các giá trị tham số tương ứng của V1 và V2 và sau đó lấy các giá trị trung bình. Tuy nhiên, đây không phải là một lựa chọn thực sự tốt vì phát âm của một phân đoạn tiếng nói không phải là độc lập, nó phụ thuộc vào các phân đoạn trước và sau nó. Để tạo cử động của môi khi phát âm tiếng Việt, luận án áp dụng mô hình của Cohhen và Massaro [23] (đã được trình bày trong

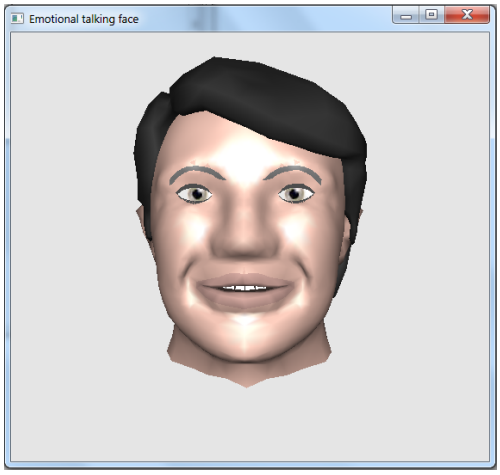
Phần 5.2) để tạo hiệu ứng đồng phát âm trên các cử động của môi khi nói. Đồng phát âm là hiệu ứng pha trộn trong đó các âm vị xung quanh sẽ có ảnh hưởng lên âm vị hiện tại.

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

Để tổng hợp cử động khuôn mặt thể hiện cảm xúc (đầu ra của mô đun EFE) và cử động của môi khi phát âm tiếng Việt, luận án áp dụng nghiên cứu được đề xuất trong [17] (đã được trình bày trong Phần 5.2). Tác giả đã đề xuất cơ chế tổng hợp các loại cử động khuôn mặt khác nhau, có khả năng tạo hoạt họa tự nhiên trên mô hình khuôn mặt 3D. Trong hệ thống khuôn mặt nói tiếng Việt, khi có xung đột xảy ra giữa các tham số ở kênh biểu cảm khuôn mặt và kênh cử động của môi khi phát âm, chúng tôi tạo quyền ưu tiên cao hơn cho cử động của môi khi phát âm tiếng 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. Cử động của môi khi phát âm và tiếng nói biểu cảm có thể được đồng bộ hóa do đầu vào của mô hình khuôn mặt đã có các mốc thời gian của từng âm vị trong tiếng nói được phát âm.

5.4 Thực nghiệm và đánh giá

Để đánh giá khả năng thể hiện cảm xúc của khuôn mặt nói tiếng Việt, chúng tôi sử dụng ParleE - môt mô hình cảm xúc cho nhân vật ảo [19], và đặt khuôn mặt trong miền cổ động viên bóng đá [16]. ParleE là một mô hình cảm xúc định lượng, linh động, và tùy biến trong đó việc đánh giá các sự kiện được dựa trên quá trình học và một giải thuật lập lịch thống kê. ParleE cũng mô hình hóa cá tính, các trạng thái thúc đẩy và vai trò của chúng trong việc quyết định cách mà nhân vật ảo trải nghiệm cảm xúc. Mô hình này được phát triển nhằm tạo cho nhân vật ảo khả năng phản ứng lại các sự kiện với các biểu cảm cảm xúc hợp lý ở các cường độ khác nhau. Chúng tôi đặt khuôn mặt nói tiếng Việt trong miền cổ động viên bóng đá [16] vì bóng đá là một trò chơi cảm xúc; có rất nhiều sự kiện trong trò chơi này kích hoạt cảm xúc không chỉ của người chơi mà cả huấn luyện viên, cổ động viên... Kiểm tra khuôn mặt trong miền cổ động viên bóng đá cho chúng ta cơ hội kiểm tra nhiều loại cảm xúc cũng như tính động



Hình 5.7: Hình ảnh minh họa video clip dùng để khảo sát cảm nhận của người dùng về cảm xúc do khuôn mặt ba chiều thể hiện.

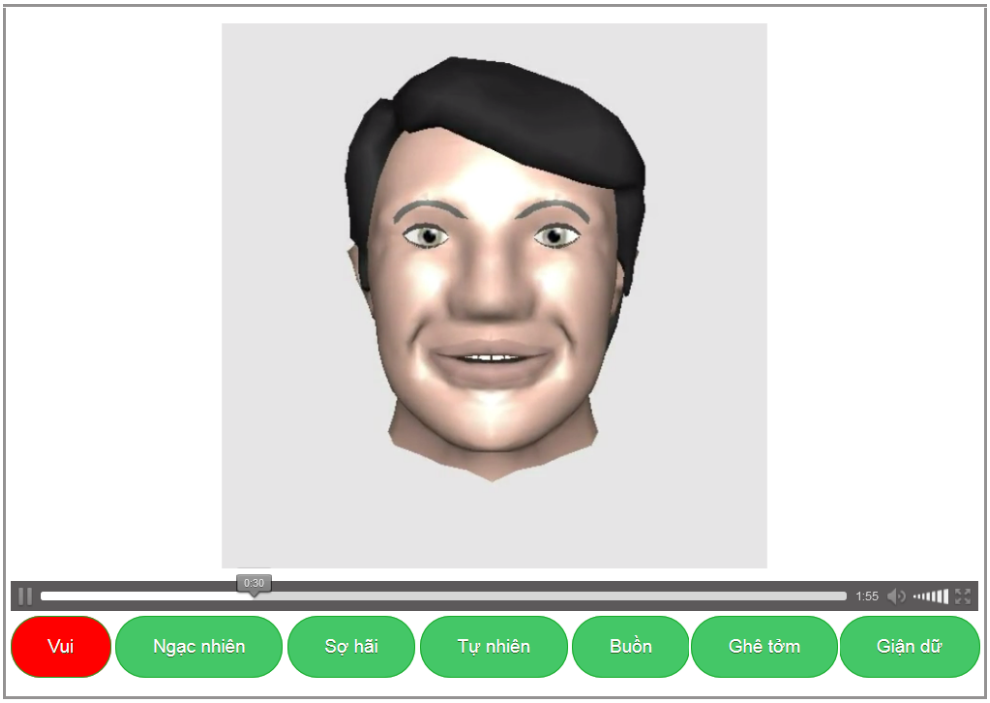
của các cảm xúc bởi vì các hành động trong một trận bóng xảy ra rất nhanh. Khuôn mặt nói tiếng Việt đóng vai trò là khuôn mặt của một cổ động viên bóng đá ảo. Nhân vật này đang xem một trận bóng đá trong đó có đội mà anh ta cổ vũ chơi. Nhân vật ảo này có thể trải nghiệm các cảm xúc khác nhau từ việc đánh giá các sự kiện dựa trên mục tiêu, tiêu chuẩn, và sở thích của anh ta. Sau đó cảm xúc sẽ được thể hiện trên khuôn mặt và trong giọng nói của khuôn mặt được xây dựng. Nói một cách ngắn gọn, mục đích của việc sử dụng ParleE và miền cổ động viên bóng đá là tạo ra đầu vào để kiểm tra, đánh giá khuôn mặt ba chiều nói tiếng Việt được xây dựng.

Chúng tôi đã tiến hành thực nghiệm để khảo sát cảm nhận của người dùng về trạng thái cảm xúc do khuôn mặt ba chiều nói tiếng Việt thể hiện. Quá trình tiến hành thực nghiệm và kết quả đánh giá như sau:

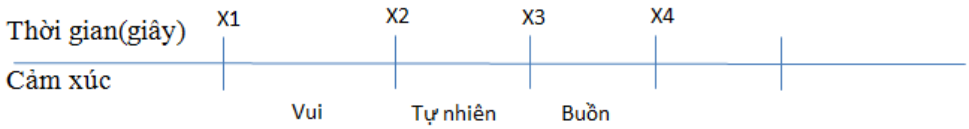
Đối tượng được đánh giá: Thực nghiệm được tiến hành với hai nhân vật ảo:

- 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 (cả hai mô đun "Tạo biểu cảm khuôn mặt" và "Tạo biểu cảm giọng điệu" đều hoạt động bình thường).

Chuẩn bị cho thực nghiệm đánh giá:



Hình 5.8: Giao diện chương trình ghi lại kết quả cảm nhận của người dùng.



Hình 5.9: Mẫu ghi kết quả cảm nhận trạng thái cảm xúc của người dùng.

Để tiến hành thực nghiệm đánh giá, chúng tôi xây dựng hai video clip cho hai nhân vật ảo A, B nói trên. Hình ảnh của video clip được minh họa trong Hình 5.7.

Mục tiêu của thực nghiệm đó là khảo sát cảm nhận của người dùng về trạng thái cảm xúc mà nhân vật ảo thể hiện. Để thực hiện mục tiêu này, chúng tôi 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.

Để ghi lại kết quả cảm nhận của người dùng, chúng tôi xây dựng một chương trình có giao diện như trong Hình 5.8. Chương trình 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

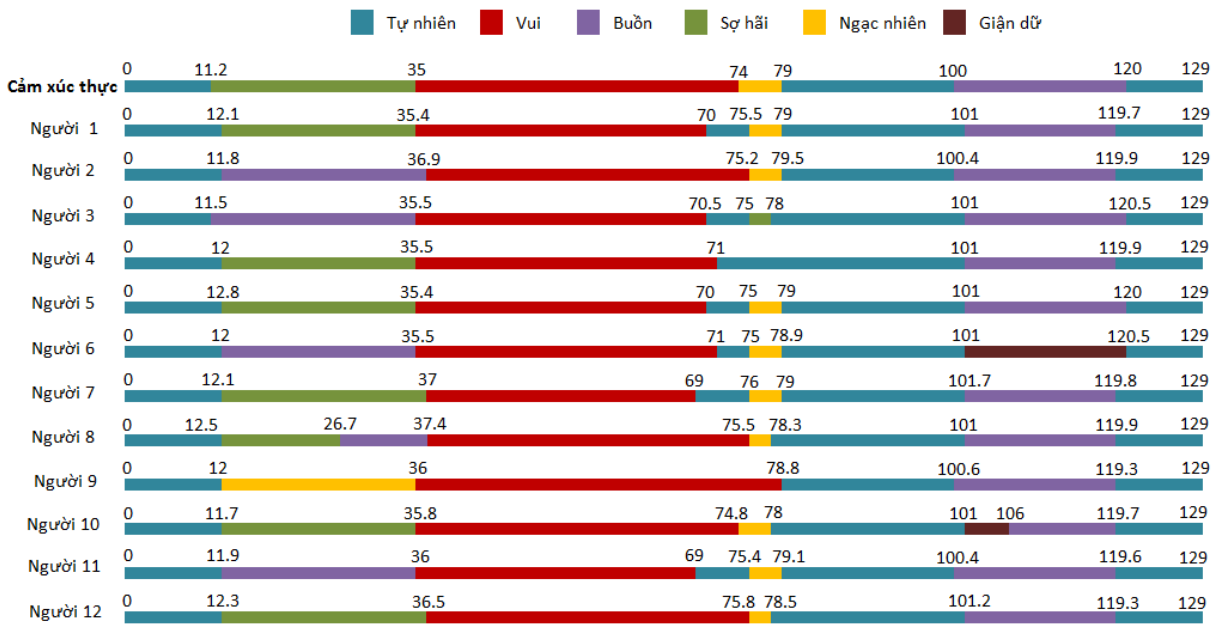
7 nút bên dưới. Ví dụ trong Hình 5.8, người dùng nhận thấy nhân vật ảo đang vui và anh ấy/cô ấy bấm vào nút "Vui". Kết quả mà chương trình trả về là các mốc thời gian (bắt đầu và kết thúc) của từng trạng thái cảm xúc mà người dùng cảm nhận được. Kết quả này có thể được ghi lại ở dạng tương tự như Hình 5.9. Các từ để mô tả cảm xúc cảm nhận được bao gồm: vui, buồn, ngạc nhiên, sợ hãi, ghê tởm, giận dữ, không cảm xúc.

Kịch bản tiến hành thực nghiệm:

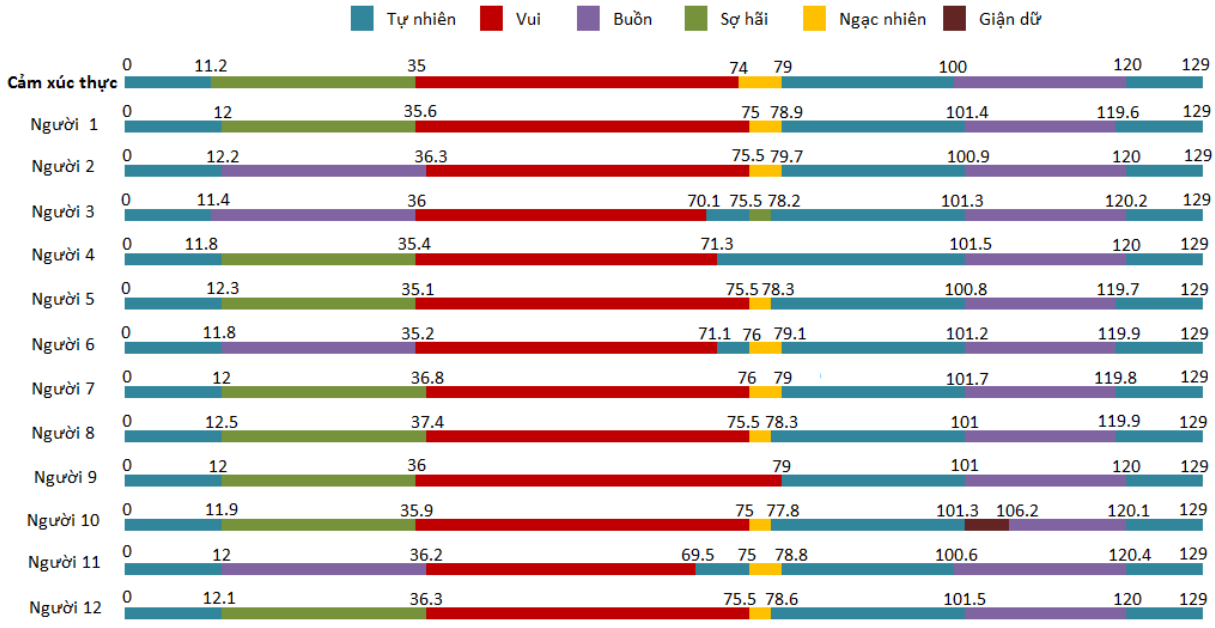
Thực nghiệm khảo sát cảm nhận của người dùng về trạng thái cảm xúc mà nhân vật ảo thể hiện được tiến hành với 12 người tham gia trong độ tuổi từ 20 đến 35, có khả năng nghe, nói, nhìn bình thường. Thực nghiệm được tiến hành trong phòng kín, cách âm tương đối tốt nhằm hạn chế tối đa ảnh hưởng của các yếu tố bên ngoài đối với kết quả đánh giá của người dùng. Mỗi phiên thực nghiệm được tiến hành riêng cho từng người như sau: Trước tiên người dùng được giới thiệu về các video clip, về mục tiêu của thực nghiệm, về chương trình ghi lại kết quả cảm nhận, người dùng cũng được hướng dẫn sử dụng chương trình này. Tiếp đến, người dùng sẽ lần lượt xem hai video clip về hai nhân vật ảo A, B đã đề cập ở trên, mỗi video clip 1 lần. Trong thời gian xem mỗi video clip, người dùng được yêu cầu ghi kết quả cảm nhận (ghi lại trạng thái cảm xúc mà họ nhận thấy nhân vật ảo đang thể hiện) bằng cách bấm vào các nút như trong Hình 5.8.

Kết quả đánh giá:

Sau khi tiến hành thực nghiệm, kết quả đánh giá của người dùng được tổng kết trong Hình 5.10 và Hình 5.11. 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 ba chiều 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 trong kết quả cảm nhận của người dùng có sự nhầm lẫn hay bỏ sót một số cảm xúc nhưng kết quả cảm nhận 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, sự sai sót đã giảm đi khá nhiều. Ví dụ, với nhân vật ảo B, sự nhầm lẫn trong việc cảm nhận



Hình 5.10: Kết quả cảm nhận của người dùng về cảm xúc do nhân vật ảo A thể hiện.



Hình 5.11: Kết quả cảm nhận của người dùng về cảm xúc do nhân vật ảo B thể hiện.

cảm xúc buồn thành cảm xúc giận dữ đã giảm đi so với 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.

5.5 Kết chương

Chương 5 của luận án đã 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. Khuôn mặt này có khả năng thể hiện cử động của môi trong khi phát âm các từ tiếng Việt một cách biểu cảm; đồng thời nó cũng có khả năng thể hiện biểu cảm khuôn mặt trong khi nói. Thực nghiệm đánh giá đã được thực hiện để kiểm tra khả năng thể hiện cảm xúc của khuôn mặt ba chiều được xây dựng. Trong thực nghiệm, khuôn mặt ba chiều đóng vai trò là khuôn mặt của một cổ động viên bóng đá ảo. Cổ động viên ảo này có thể trải nghiệm các cảm xúc khác nhau, từ đó thể hiện các biểu cảm trên khuôn mặt và trong giọng nói. Kết quả thực nghiệm cho thấy khuôn mặt ba chiều được xây dựng có khả năng thể hiện cảm xúc khá tốt.

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 nghiệm 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 không cảm xúc được xây dựng. Sau đó, các luật này được sử dụng để biến đổi tiếng nói tiếng Việt ở trạng thái không cảm xúc 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.* Proc. 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.* Proc. 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.* Proc. 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 ana-*

lyzing 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 Talking Face for Embodied Conversational Agents.* In Proc. of the 11th IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF 2015), pp.94-99.

TÀI LIỆU THAM KHẢO

[1] X. T. Đỗ and H. T. Lê. *Giáo trình tiếng Việt 2.* Nhà xuất bản đại học Sư Phạm, 2007.

[2] D. Abercrombie. *Elements of general phonetics.* Chicago: Alding, 1967.

[3] J. Ahlberg. Candide-3 – an updated parameterized face. Technical Report Report No. LiTH-ISY-R-2326, Dept. of Electrical Engineering, Linkoping University, Sweden, 2001.

[4] I. Albrecht. *-Faces and Hands- Modeling and animating anatomical and photorealistic models with regard to the communicative competence of virtual humans.* PhD thesis, University at des Saarlandes, 2005.

[5] I. Albrecht, J. Haber, K. Kähler, M. Schröder, and H. P. Seidel. May i talk to you? :-) facial animation from text. In *Proceedings Pacific Graphics 2002*, pages 77–86, 2002.

[6] F. H. Allport. *Social psychology.* Houghton Mifflin, Boston, 1924.

[7] R. J. Andrews. The information potentially avaiable in mammal displays. *Non-verbal communication*, 1972.

[8] M. B. Arnold. *Emotion and personality: Vol. 1(2). Psychological aspects.* Columbia University Press, New York, 1960.

[9] J. R. Averill. A constructivist view of emotion. *Emotion: Theory, research and experience*, I:305–339, 1980.

[10] R. Barra-Chicote, J. Yamagishi, S. King, J. M Montero, and J. Macias Guarasa. Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. *Speech Communication*, 52(5):394–404, 2010.

[11] J. Bates. The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125, July 1994.

[12] E. G. Beier and A. Zautra. Identification of vocal communication of emotion across cultures. *Journal of Consulting and Clinical Psychology*, 1972.

[13] R. L. Birdwhistell. *Kinesics and context.* Philadelphia: University of Pennsylvania Press, 1970.

[14] D. Buhler and H. Hetzer. *Testing children's development from birth to school age.* New York: Farrar & Rinehart, 1935.

[15] T. D. Bui, D. Heylen, and A. Nijholt. Improvements on a simple muscle-based 3d face for realistic facial expressions. In *16th Int. Conf. on Computer Animation and Social Agents*, pages 33–40. IEEE Computer Society, 2003.

[16] T. D. Bui, D. Heylen, and A. Nijholt. Building embodied agents that experience and express emotions: A football supporter as an example. In *Proc. CASA2004.* Computer Graphics Society, 2004.

[17] T. D. Bui, D. Heylen, and A. Nijholt. Combination of facial movements on a 3d talking head. In *Proc. CGI2004.* IEEE Computer Society, 2004.

[18] T. D. Bui, D. Heylen, M. Poel, and A. Nijholt. Generation of facial expressions from emotion using a fuzzy rule based system. In *Australian Joint Conf. on Artificial Intelligence (AI 2001)*, pages 83–95, Berlin, 2001. Lecture Notes in Computer Science, Springer.

[19] T. D. Bui, D. Heylen, M. Poel, and A. Nijholt. Parlee: An adaptive plan-based event appraisal model of emotions. In *KI 2002: Advances in Artificial Intelligence*, pages 129–143, Berlin, 2002. Lecture Notes in Computer Science, Springer.

[20] F. Burkhardt. Emofilt: the simulation of emotional speech by prosody-transformation. *Proc. of Interspeech*, 2005.

[21] J. P. Cabral and L. C. Oliveira. Emo voice: a systemto generate emotions in speech. *In Proc. INTERSPEECH*, 2006.

[22] J. E. Cahn. The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, pages 1–19, 1990.

[23] M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. In *Models and Techniques in Computer Animation*, pages 139–156.

[24] Gary Collier. *Emotional expression.* Lawrence Erlbaum Associates, New Jersey, 1985.

[25] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[26] R. H. Cornelius. *The Science of Emotion.* Upper Saddle River, NJ, 1996.

[27] A. R. Damasio. *Descartes' error: Emotion, reason, and the human brain.* G.P. Putnam, New York, 1994.

[28] C. Darwin. *The expression of the emotions in man and animals.* Univerity of Chicago Press, Chicago, 1872/1965.

[29] D. C. DeCarlo, M. S. Revilla, and J. Venditti. Making discourse visible: Coding and animating conversational facial displays. In *Computer Animation 2002*, 2002.

[30] Shichuan Du, Yong Tao, and Aleix M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences. Edited by David J. Heeger, New York University, New York, NY*, 2014.

[31] P. Ekman. Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation 1971*, 19, 1972.

[32] P. Ekman. *Darwin and facial expression: A century of research in review.* Academic Press, New York, 1973.

[33] P. Ekman. Biological and cultural contributions to body and facial movement. In J. Blacking, editor, *The anthropology of the body.* Academic Press, London, 1977.

[34] P. Ekman. *Emotion in the human face.* Cambridge University Press, Cambridge, 1982.

[35] P. Ekman. Expression and the nature of emotion. In K.R. Scherer and P. Ekman, editors, *Approaches to Emotion.* Lawrence Erlbaum, Hillsdale, NJ, 1984.

[36] P. Ekman and W. V. Friesen. *Unmasking the Face: A Guide To Recognizing Emotions From Facial Clues.* Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

[37] P. Ekman and W. V. Friesen. *Facial Action Coding System.* Consulting Psychologists Press, Palo Alto, CA, 1978.

[38] P. Ekman and J. Hager. Facial action coding system affect interpretation database (facsaid). *Retrieved from http://face-and-emotion.com/dataface/facsaid/description.jsp*, 2002.

[39] P. Ekman, R. W. Levenson, and W. V. Friesen. Autonomic nervous system activity distinguishes between emotions. *Science*, 221, 1983.

[40] P. Ekman and E. L. Rosenberg. *What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS).* Illustrated Edition, Oxford University Press, 1997.

[41] M. S. El-Nasr, J. Y., and T. R. Ioerger. FLAME-fuzzy logic adaptive model of emotions. *Autonomous Agents and Multi-Agent Systems*, 3(3):219–257, 2000.

[42] C. Elliott. *The Affective Reasoner: A Process model of emotions in a multi-agent system.* PhD thesis, Northwestern University, Evanston, IL, 1992.

[43] D. Erickson. Expressive speech: Production, perception and application to speech synthesis. *Acoust. Sci. & Tech*, 26:317–325, 2005.

[44] I. A. Essa and A. Pentland. A vision system for observing and extracting facial action parameters. In *Proceedings of IEEE Computer Vision and Pattern Recognition Conference*, 1994.

[45] R. Fernandez and B. Ramabhadran. Automatic exploration of corpus specific properties for expressive text-to-speech: a case study in emphasis. *In Proc. ISCA workshop on speech synthesis*, pages 34–39, 2007.

[46] J. M. Fernández-Dols and J. M. Carroll. Is the meaning perceived in facail expression independent of its context? In J. A. Russell and J. M. Fernández-Dols, editors, *The Psychology of Facial Expression*. Cambridge University Press, New York, NY, 1997.

[47] V. C. Flores. Artnatomy (anatomical basis of facial expression interactive learning tool). In *Proceedings of the ACM Educators Program (SIGGRAPH 06), New York, NY, USA*, 2006.

[48] J. P. Forgas and S. Moylan. After the movies: the effects of transient mood states on social judgments. *Personality and Social Psychology Bulletin*, 13, 1987.

[49] A. J. Fridlund. *Human facial expression: An evolutionary view.* Academic, New York, 1994.

[50] W. Friesen and P. Ekman. *EMFACS-7: Emotional Facial Action Coding System.* Unpublished manual, University of California, California, 1983.

[51] N. H. Frijda. *The emotions.* Cambridge University Press, Cambridge, 1986.

[52] D. H. Galernter. *The muse in the machine.* Free Press, New York, 1994.

[53] D. Govind and S. R. Mahadeva Prasanna. Expressive speech synthesis: a review. *International Journal of Speech Technology*, 16:237–260, 2012.

[54] J. Gratch. Émile: Marshalling passions in training and education. In M. Gini C. Sierra and J. S. Rosenschein, editors, *Proceedings of the Fourth International Conference on Autonomous Agents*, pages 325–332, Barcelona, Catalonia, Spain, 2000. ACM Press.

[55] W. W. Grings and M. E. Dawson. *Emotions and bodily responses: a psychophysiological approach.* Academic Press, New York, 1978.

[56] J. Hager and P. Ekman. Essential behavioral science of the face and gesture that computer scientists need to know. In *Proceedings of the International Workshop on Automatic Face and Gesture Recognition*, 1995.

[57] H. Hanson. Glottal characteristics of female speakers: acoustic correlates. *J. Acoust. Soc. Am. 101.*, pages 466–481, 1997.

[58] B. Hayes-Roth and R. van Gent. Story-making with improvisational puppets. In W. L. Johnson and B. Hayes-Roth, editors, *Proceedings of the 1st International Conference on Autonomous Agents*, pages 1–7, New York, 1997. ACM Press.

[59] D. Heylen, M. Theune, R. op den Akker, and A. Nijholt. Social agents: The first generations. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, pages 1–7, 2009.

[60] G. Hofer, K. Richmond, and R. Clark. Informed blending of databases for emotional speech synthesis. *In Proc. INTERSPEECH*, 2005.

[61] C. F. Huang and M. Akagi. A rule-based speech morphing for verifying an expressive speech perception model. *Proc. Interspeech*, pages 2661–2664, 2007.

[62] G. L. Huttar. Relations between prosodic variables and emotions in normal american english utterances. *Journal of Speech and Hearing Research*, 11:481–487, 1968.

[63] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura. A speech synthesis system for assisting communications. *In ISCA workshop on speech and emotion*, pages 167–172, 2000.

[64] Z. Inanoglu and S. Young. A system for transforming the emotion in speech: Combining data-driven conversion techniques for prosody and voice quality. *Proc. of Interspeech*, 2007.

[65] John Ingram and Thu Nguyen. Stress, tone and word prosody in vietnamese compounds. *Proceedings of the 11th Australian International Conference on Speech Science & Technology*, pages 193–198, 2006.

[66] K. Inoue, K. Kawabata, and H. Kobayashi. On a decision making system with emotion. In *IEEE Int. Workshop on Robot and Human Communication*, 1996.

[67] K. Isbister and P. Doyle. Design and evaluation of embodied conversational agents: a proposed taxonomy. In *In Proceedings of AAMAS 2002 Workshop on Embodied Conversational Agents ? Let?s Specify and Evaluate Them!, Bologna, Italy*, 2002.

[68] C. T. Ishii and N. Campbell. Analysis of acoustic-prosodic features of spontaneous expressive speech. *Proceedings of 1st International Congress of Phonetics and Phonology, 19*, 2002.

[69] C. Izard. Emotions and facial expressions: A perspective from differential emotions theory. *In Russell, J. and Fernandez-Dols, J., editors, The Psychology of Facial Expression. Maison des Sciences de l'Homme and Cambridge University Press*, 1997.

[70] C. E. Izard. *The face of emotion.* Appleton-Century-Crofts, New York, 1971.

[71] C. E. Izard. Differential emotions theory and the facial feedback hypothesis activation: Comments on tourangeau and ellsworth's 'the role of facial response in experience of emotion'. *Journal of Personality and Social Psychology*, 40:350–354, 1981.

[72] C. E. Izard. The substrates and functions of emotion feelings: William james and current emotion theory. *Personality and Social Psychology Bulletin*, 16(4):626–635, 1990.

[73] C. E. Izard. Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin*, 115(2):288–299, 1994.

[74] W. James. What is an emotion. *Mind*, 19:188–205, 1884.

[75] D. Jay. Field studies. *Behavior of nonhuman primates: Modern reserach trends*, 1, 1965.

[76] A. Kappas. What facial activity can and cannot tell us about emotions. In M. Katsikitis, editor, *The human face: Measurement and meaning*, pages 215–234. Kluwer Academic Publishers, Dordrecht, 2003.

[77] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne. Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sounds. *Speech Communication*, 27:187–207, 1999.

[78] R. D. Kent and C. Read. *Acoustic Analysis of Speech*. San Diego: Singular Publishing Group, 1992.

[79] S. A. King, R. E. Parent, and B. Olsafsky. An anatomically-based 3d parametric lip model to support facial animation and synchronized speech. In *Proceedings of Deform 2000*, pages 7–19, 2000.

[80] S. Kshirsagar and N. Magnenat-Thalmann. A multilayer personality model. In *Proceedings of 2nd International Symposium on Smart Graphics*, pages 107–115. ACM Press, 2002.

[81] D. Kurlander, T. Skelly, and D. Salesin. Comic chat. In *SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 225–236, 1996.

[82] J. D. Laird and C. Bresler. William james and the mechanisms of emotional experience. *Personality and Social Psychology Bulletin*, 16, 1990.

[83] C. Latta, N. Alvarado, S. S. Adams, and S. Burbeck. An expressive system for animating characters or endowing robots with affective displays. In *Society for Artificial Intelligence and Social Behavior (AISB), 2002 Annual Conference, Symposium on Animating Expressive Characters for Social Interactions*, 2002.

[84] R. S. Lazarus. Progress on a cognitive-motivational-relational theory ofs emotion. *American Psychologist*, 46, 1991.

[85] R. S. Lazarus and E. Alfert. Short circuiting of threat by experimentally altering cognitive appraisal. *Journal of Abnormal and Social Psychology*, 69, 1964.

[86] R. S. Lazarus, I. R. Averill, and E. M. Opton. Feeling and emotion. In M. B. Arnold, editor, *Toward a Cognitive Theory of Emotion*, pages 207–232. Academic Press, New York, 1970.

[87] Hong Minh Le and Khanh Hung Le. Analysis and synthesis for duration feature of vietnamese. *The 6th National Conference in Information Technology, Thainguyen, Vietnam*, 2003.

[88] Hong Minh Le and Tuan Ngoc Quach. Some results in phonetic analysis to vietnamese text-to-speech synthesis based on rules. *Journal on Information and Communication Technology*, 2006.

[89] Tang Ho Le, Anh Viet. Nguyen, Vinh Hao. Truong, Van Hien. Bui, and Dung Le. A study on vietnamese prosody. *New Challenges for Intelligent Information and Database Systems*, 351:63–73, 2011.

[90] L. Leinonen. Expression of emotional-motivational connotations with a one-word utterance. *J. Acoust. Soc. Am.*, 102:1853–1863, 1997.

[91] R. W. Levenson. Autonomic nervous system differences among emotions. *Psychological Science*, 3:23–27, 1992.

[92] C. L. Lisetti. Emotion generation for artificial agents via a hybrid architecture. In *Proceedings of the Autonomous Agents Workshop on Emotion-Based Agent Architectures (EBAA'99)*, 1999.

[93] Dang Khoa Mac, Eric Castelli, Véronique Aubergé, and Albert Rilliard. How vietnamese attitudes can be recognized and confused: Cross-cultural perception and speech prosody analysis. *International Conference on Asian Language Processing*, pages 220–223, 2011.

[94] K. Maekawa. Phonetic and phonological characteristics of paralinguistic information in spoken japanese. *Proc. Int. Conf. Spoken Language Processing*, pages 635–638, 1998.

[95] C. Mahardika, R. Itimad, H. Ahmad, and S. Nadz. Eye, lip and crying expression for virtual human. *International Journal of Interactive Digital Media*, 1(2), 2013.

[96] C. Z. Malatesta. Infant emotion and the vocal effect lexicon. *Motivation and Emotion*, 5(1):1–23, 1981.

[97] G. Mandler. *Mind and Emotion*. Wiley, New York, 1975.

[98] A. Mehrabian. Communication without words. *Psychology Today*, 2(4):53–56, 1968.

[99] A. Mehrabian. *Nonverbal communication*. Chicago: Aldine-Atherson, 1972.

[100] C. Menezes, K. Maekawa, and H. Kawahara. Perception of voice quality in pralinguistic information types: A preliminary study. *Proceedings of the 20th General Meeting of the PSJ*, pages 153–158, 2006.

[101] K. Miyanaga, T. Masuko, and T. Kobayashi. A style control techniques for hmm-based speech synthesis. *In Proc. ICSLP*, 2004.

[102] S. Mohammad Mavadati, H. Mahoor Mohammad, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[103] J. M. Montero, J. Gutierrez-Arriola, J. Colas, E. Enriquez, and J. M. Pardo. Analysis and modelling of emotional speech in spanish. *In Proc. ICPhS*, pages 671–674, 1999.

[104] I. R. Murray and J. L. Arnott. Toward the simulation of emo- tion in synthetic speech: a review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93:1097–1108, 2006.

[105] T. K. Nguyen, D. T. Nguyen, T. H. Le, and V. L. Trinh. Dsp-based embedded system for text-to-speech synthesis of vietnamese. In *Proceedings of the 2nd Asia Pacific International Conference on Information Science and Technology.*, pages 215–219, 2007.

[106] T. L. Nguyễn and T. H. Nguyễn. *Tiếng Việt (Ngữ âm và Phong cách học).* Nhà xuất bản đại học Sư Phạm, 2007.

[107] T. Nose, J. Yamagishi, and T. Kobayashi. A style control technique for hmm-based expressive speech synthesis. *IEICE Transactions on Information and Systems E*, 90-D(9):1406–1413, 2007.

[108] K. Oatley and P. N. Johnson-Laird. Towards a cognitive theory of emotions. *Cognition and Emotion*, 1(1):29–50, 1987.

[109] J. J. Ohala. The nonlinguistic components of speech. *Speech evaluation in psychiatry*, pages 39–49, 1981.

[110] A. Ortony, G. L. Clore, and A. Collins. *The Cognitive Structure of Emotions.* Cambridge University Press, Cambridge, England., 1988.

[111] M. Pantic and I. Patras. Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. *Proc. IEEE conf. Systems, Man and Cybernetics*, 4:3358–3363, 2005.

[112] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. *Proc. 13th ACM Int'l Conf. Multimedia and Expo*, pages 317–321, 2005.

[113] M. D. Pell. Influence of emotion and focus location on prosody in matched statements and questions. *J. Acoust. Soc. Am.*, 109:1668–1680, 2001.

[114] K. Perlin and A. Goldberg. Improv: A system for scripting interactive actors in virtual worlds. *Computer Graphics*, 30(Annual Conference Series):205–216, 1996.

[115] R. Picard. *Affective Computing.* MIT Press, Cambridge, MA, 1997.

[116] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny. The ibm expressive text to speech synthesis system for american english. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:1099–1109, 2006.

[117] R. Plutchik. Emotions: A general psychoevolutionary theory. In K. R. Scherer and P. Ekman, editors, *Approaches to emotion.* Lawrence Erlbaum, London, 1984.

[118] A. Raouzaiou, K. Karpouzis, and S. D. Kollias. Online gaming and emotion representation. In N. N. García, J. M. Martínez, and L. Salgado, editors, *Volume 2849 of Lecture Notes in Computer Science*, pages 298–305. Springer, 2003.

[119] W. S. Reilly. Believable social and emotional agents. Technical Report Ph.D. Thesis. Technical Report CMU-CS-96-138, Carnegie Mellon University, Pittsburgh, PA, USA, 1996.

[120] W. S. Reilly and J. Bates. Building emotional agents. Technical Report CMU-CS-92-143, Carnegie Mellon University, Pittsburgh, PA, USA, 1992.

[121] J. Rickel and W. L. Johnson. Steve: A pedagogical agent for virtual reality. In *Proceedings of the Second International Conference on Autonomous Agents*, 1998.

[122] I. J. Roseman. Cognitive determinants of emotions: A structural theory. *Review of Personality and Social Psychology*, 5, 1984.

[123] R. Rosenthal, J. A. Hall, M. R. DiMatteo, P. L. Rogers, and D. Archer. *Sensitivity to nonverbal communication: The PONS test.* Baltimore: John Hopkins University Press, 1979.

[124] J. A. Russell and J. M. Fernández-Dols. What does a facial expression mean? *In Russell, J. A. and Fernández-Dols, J. M., editors, The Psychology of Facial Expression. Cambridge University Press, New York, NY*, 1997.

[125] T. Saitou, M. Goto, M. Unoku, , and M. Akagi. Speech-to-singing synthesis: converting speaking voices to singing voices by controlling acoustic features unique to singing voices. *Proc. WASPAA2007*, 2007.

[126] K. R. Scherer. Nonlinguistic vocal indicators of emotion and psychopathology. *Emotion in personality and psychopathology*, pages 493–529, 1979.

[127] K. R. Scherer. What does facial expression express? In K. Strongman, editor, *International Review of Studies on Emotion*, volume 2. Wiley, Chichester, 1992.

[128] K. R. Scherer. Appraisal considered as a process of multilevel sequential checking. In K. Scherer, A. Schorr, and T. Johnstone, editors, *Appraisal processes in emotion.* Oxford University Press, Oxford, 2001.

[129] K. R. Scherer. Vocal communication of emotion: a review of research paradigms. *Speech Communication*, 40:227–256, 2003.

[130] K. R. Scherer, R. Banse, H. G. Wallbott, and T. Goldbeck. Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, 15:123–148, 1991.

[131] M. Schroder. Expressive speech synthesis: past, present and possible futures. *Affective Information Processing*, 2:111–126, 2009.

[132] P. R. Shaver, S. Wu, and J. C. Schwartz. Cross-cultural similarities and differences in emotion and its representation: A prototype approach. *Review of Personality & Social Psychology*, 13, 1992.

[133] S. A. Shields and R. M. Stern. Emotion: The perception of bodily change. In P. Pliner, K. R. Blankstein, and I. M. Spigel, editors, *Perception of emotion in self and others*, pages 85–106. Plenum, New York, NY, 1979.

[134] K. Sjölander and J. Beskow. Wavesurfer - an open source speech tool. In *Proceedings of the 6th International Conference of Spoken Language Processing (ICSLP)*, 2000.

[135] C. A. Smith. Dimensions of appraisal and physiological response in emotion. *Journal of Personality and Social Psychology*, 56:339–353, 1989.

[136] J. C. Speisman, R. S. Lazarus, A. Mordkoff, and L. Davison. Experimental reduction of stress based on ego-defensive theory. *Journal of Abnormal and Social Psychology*, 68:367–380, 1964.

[137] A. Stern, A. Frank, , and B. Resner. Virtual petz: A hybrid approach to creating autonomous, lifelike dogz and catz. In K. P. Sycara and M. Wooldridge, editors, *Proceedings of the 2nd International Conference on Autonomous Agents (Agents'98)*, pages 334–335, New York, 1998. ACM Press.

[138] EAR. Tanguy. *Emotions: the Art of Communication Applied to Virtual Actors*. PhD thesis, Universit of Bath, 2006.

[139] J. Tao, Y. Kang, and A. Li. Prosody conversion from neutral speech to emotional speech. *IEEE Trans. on Audio, Speech and Language Processing*, 14:1–19, 2007.

[140] M. Theune, S. Faas, D. Heylen, and A. Nijholt. The virtual storyteller: Story creation by intelligent agents. In *Proceedings TIDSE 03: Technologies for Interactive Digital Storytelling and Entertainment*, pages 204–215. Fraunhofer IRB Verlag, 2003.

[141] M. Theune, K. Meijs, D. Heylen, and R. Ordelman. Ieee transactions on audio, speech, and language processing. *SAffective Information Processing*, 14(4):1099–1108, 2006.

[142] F. Thomas and O. Johnston. *The Illusion of Life*. Abbeville Press, New York, 1981.

[143] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence.*, 23(2):97–115, 2001.

[144] S. S. Tomkins. *Affect, Imagery, Consciousness (Volume 1): The Positive Affects*. Springer, New York, 1962.

[145] S. S. Tomkins. *Affect, Imagery, Consciousness (Volume 1): The Negative Affects*. Springer, New York, 1963.

[146] Do Dat Tran, Eric Castelli, Jean-François Serignat, and Viet Bac Le. Analysis and modeling of syllable duration for vietnamese speech synthesis. *O-COCOSDA*, 2007.

[147] J. D. Velásquez. Modeling emotions and other motivations in synthetic agents. In *Proceedings of the 14th National Conference on Artificial Intelligence and 9th Innovative Applications of Artificial Intelligence Conference (AAAI-97/IAAI-97)*, pages 10–15, Menlo Park, 1997. AAAI Press.

[148] P. Viola and M. Jones. Robust real-time object detection,. *Tech. rep.,Cambridge Research Laboratory Technical report series.*, (2), 2001.

[149] J. Vroomen, R. Collier, and S. J. L. Mozziconacci. Duration and intonation in emotional speech. *In Proc. EUROSPEECH*, pages 577–580, 1993.

[150] Q Vu and et al. Vos: The corpus-based vietnamese text-to-speech system. *Journal on Information, Technologies, and Communications*, 2010.

[151] T.T Vu, C.M Luong, and S. Nakamura. An hmm-based vietnamese speech synthesis system. In *Proceedings of the 12th International Oriental COCOSDA Conference.*, pages 116–121, 2009.

[152] F. Wallhoff. The facial expressions and emotions database homepage (feedtum). *www.mmk.ei.tum.de/ waf/fgnet/feedtum.html*, 2005.

[153] M. Wooldridge. Intelligent agents. In G. Weiss, editor, *Multiagent Systems. A Modern Approach to Distributed Artificial Intelligence*. Cambridge, Mass, MIT Press, 1999.

[154] J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano. Model adaptation approach to speech synthesis with diverse voices and styles. *In Proc. ICASSP*, pages 1233–1236, 2007.

[155] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi. Modeling of various speaking styles and emotions for hmm-based speech synthesis. *In Proc. EUROSPEECH*, pages 2461–2464, 2003.

[156] S. Zhang, Z. Wu, H.M. Meng, and L. Cai. Facial expression synthesis based on emotion dimensions for affective talking avatar. In *Modeling Machine Emotions for Realizing Intelligence, SIST*, pages 109–132. Springer-Verlag Berlin Heidelberg., 2010.

PHỤ LỤC 1
Hệ mã hóa các cử động khuôn mặt

Hệ thống mã hóa các cử động khuôn mặt (Facial Action Coding System - FACS) được đề xuất bởi Paul Ekman và Wallace Friesen [37]. Hệ thống này mô tả tất cả các cử động cơ bản có thể quan sát được của khuôn mặt. FACS là một danh sách gồm 64 đơn vị cử động (tên tiếng Anh là Action Unit, viết tắt là AU), mỗi AU được mô tả là kết hợp của một hoặc một số các cơ trên khuôn mặt. Bảng 5.1 mô tả danh sách các AU của FACS.

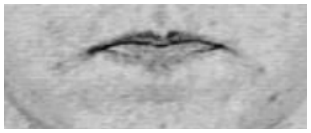
Bảng 5.1: Hệ mã hóa các cử động khuôn mặt (FACS).

AU	**Mô tả chuyển động**	**Các cơ chuyển động**	**Hình ảnh minh họa**
1	Nhướn mày trong	Frontalis, pars medialis	
2	Nhướn mày ngoài	Frontalis, pars lateralis	
4	Hạ lông mày	Corrugator supercilii, Depressor supercilii	
5	Nhướn mi trên	Levator palpebrae superioris	
6	Nâng má	Orbicularis oculi, pars orbitalis	
7	Căng mí mắt	Orbicularis oculi, pars palpebralis	

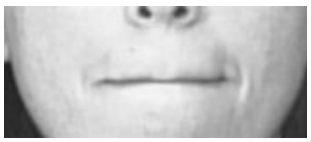
Bảng 5.1 Tiếp tục từ trang trước

AU	Mô tả chuyển động	Các cơ chuyển động	Hình ảnh minh họa
9	Nhăn mũi	Levator labii superioris alaquae nasi	
10	Nâng môi trên	Levator labii superioris	
11	Làm sâu mũi	Levator anguli oris (a.k.a. Caninus)	
12	Kéo khóe môi	Zygomaticus major	
13	Phồng má	Zygomaticus minor	
14	Má lúm đồng tiền	Buccinator	
15	Nén khóe môi	Depressor anguli oris (a.k.a. Triangularis)	
16	Bặm môi dưới	Depressor labii inferioris	

Bảng 5.1 Tiếp tục từ trang trước

AU	Mô tả chuyển động	Các cơ chuyển động	Hình ảnh minh họa
17	Nâng cằm	Mentalis	
18	Nhàu môi	Incisivii labii superioris and Incisivii labii inferioris	
20	Kéo căng môi	Risorius w/ platysma	
22	Môi hình phễu	Orbicularis oris	
23	Bặm chặt môi	Orbicularis oris	
24	Ép môi	Orbicularis oris	
25	Tách môi trên và dưới	Depressor labii inferioris or relaxation of Mentalis, or Orbicularis oris	
26	Hạ hàm	Masseter, relaxed Temporalis and internal Pterygoid	

Bảng 5.1 Tiếp tục từ trang trước

AU	Mô tả chuyển động	Các cơ chuyển động	Hình ảnh minh họa
27	Căng miệng	Pterygoids, Digastric	
28	Mút môi	Orbicularis oris	
41	Rủ mí mắt	Relaxation of Levator palpebrae superioris	
42	Ti hí mắt	Orbicularis oculi	
43	Nhắm mắt	Relaxation of Levator palpebrae superioris; Orbicularis oculi, pars palpebralis	
44	Liếc mắt	Orbicularis oculi, pars palpebralis	
45	Chớp mắt	Relaxation of Levator palpebrae superioris; Orbicularis oculi, pars palpebralis	
46	Nháy mắt	Relaxation of Levator palpebrae superioris; Orbicularis oculi, pars palpebralis	

Bảng 5.1 Tiếp tục từ trang trước

AU	Mô tả chuyển động	Các cơ chuyển động	Hình ảnh minh họa
51	Quay đầu sang trái		
52	Quay đầu sang phải		
53	Ngửa đầu lên		
54	Cúi đầu xuống		
55	Đầu ngả sang trái		
56	Đầu ngả sang phải		
57	Ngả đầu về phía trước		

Bảng 5.1 Tiếp tục từ trang trước

AU	Mô tả chuyển động	Các cơ chuyển động	Hình ảnh minh họa
58	Ngả đầu về phía sau		
61	Liếc mắt sang trái		
62	Liếc mắt sang phải		
63	Ngước mắt lên trên		
64	Mắt nhìn xuống dưới		

PHỤ LỤC 2
Cơ sở dữ liệu tiếng Việt có cảm xúc

Cơ sở dữ liệu được sử dụng trong Chương 4 bao gồm các phát âm của 19 câu sau:

1. Có thư mới.
2. Làm gì có chuyện mà tôi tức điên lên.
3. Nghe đâu điểm hẹn là Huế thì phải.
4. Tôi đã mua ô tô mới.
5. Bỏ mấy cái thư không dùng nữa đi.
6. Mê tín, cổ hủ đến như thế.
7. Tôi đã nhận được sự cổ vũ của mọi người.
8. Chắc là thư đã đến nơi.
9. Tôi theo dõi liên tục.
10. Đến chỗ tôi.
11. Xin chân thành cảm ơn.
12. Xin thành thật xin lỗi.
13. Tôi sẽ không nói cảm ơn đâu.
14. Tôi cảm thấy chóng mặt.
15. Tôi cũng đã sai.
16. Có cần chỗ ngồi để xem pháo hoa không?
17. Tôi chẳng nói rồi còn gì, nếu không làm thì...

18. Hãy nói cho tôi biết lý do đến không đúng giờ.

19. Sẽ tập trung lại ở chỗ nghỉ chân nhé.

PHỤ LỤC 3
Kiểm định giả thuyết thống kê

Bài toán kiểm định giả thuyết thống kê là một bài toán lớn và quan trọng của thống kê toán học. Trong phụ lục này, chúng tôi sẽ đề cập đến một số định nghĩa và vấn đề liên quan đến bài toán kiểm định giả thuyết thống kê.

Một số định nghĩa

Giả thuyết

Một mệnh đề (một câu khẳng định) về một vấn đề chưa biết nào đó được gọi là một giả thuyết. Ta thường dùng H_0 để chỉ một giả thuyết. Giả thuyết là một mệnh đề có thể đúng hoặc không đúng.

Đối thuyết

Một mệnh đề trái với giả thuyết được gọi là đối thuyết. Ta thường dùng H_1 để chỉ đối thuyết.

Kiểm định giả thuyết thống kê

Một phép kiểm định (hay trắc nghiệm) một giả thuyết thống kê là một qui tắc, theo đó, dựa vào một mẫu cụ thể được thực hiện, chúng ta có thể quyết định chấp nhận hay bác bỏ giả thuyết đang xét.

Kiểm định một giả thuyết thống kê không phải là một phép chứng minh về tính đúng hoặc không đúng của giả thuyết. Kiểm định một giả thuyết thống kê thực chất là xây dựng một qui tắc hành động dựa vào mẫu đã có đưa ra quyết định lựa chọn giả thuyết H_0 hoặc đối thuyết H_1.

Sai lầm

Nếu chúng ta bác bỏ giả thuyết H_0 khi, thực ra, nó phải được chấp nhận thì chúng ta đã mắc phải sai lầm gọi là **Sai lầm loại I**. Nếu chúng ta chấp nhận H_0 trong khi, thực ra, nó phải bị bác bỏ thì chúng ta đã mắc phải sai lầm gọi là **Sai lầm loại II**.

Xác suất mắc phải sai lầm loại I, thường ký hiệu là α, gọi là **Mức ý nghĩa** của kiểm định. Như vậy, xác suất để chấp nhận H_0 khi nó đúng là $(1 - \alpha)$.

Nếu ký hiệu β là xác suất mắc phải sai lầm loại II, thì xác suất để bác bỏ H_0 khi nó sai là $(1 - \beta)$, được gọi là **Năng lực của phép kiểm định**.

Như vậy, một báo cáo kết quả so sánh là "sự khác biệt có ý nghĩa về mặt thống kê ở mức ý nghĩa 5%" có nghĩa là "giả thuyết không" H_0 đã bị bác bỏ với nguy cơ sai lầm là 5%.

Các bước cơ bản của một phép kiểm định giả thuyết thống kê

1. Đặt giả thuyết không H_0 và đối thuyết H_1
2. Xác định mức ý nghĩa α của phép kiểm định
3. Với cặp giả thuyết và mức ý nghĩa α đã xác định, chúng ta thiết lập được một *Qui luật quyết định* dùng để quyết định chấp nhận hay bác bỏ giả thuyết H_0. Qui luật này bao gồm việc chọn một thống kê thích hợp để dùng cho phép kiểm định và đưa ra một giá trị tới hạn để so sánh.

Khác với phép kiểm định một giả thuyết với mức ý nghĩa α cho trước, các nhà nghiên cứu thường xác định mức ý nghĩa nhỏ nhất, tại đó "giả thuyết không" H_0 bị bác bỏ. Từ đó, người ta có định nghĩa: Trong một phép kiểm định, mức ý nghĩa nhỏ nhất, tại đó "giả thuyết không" H_0 có thể bị bác bỏ được gọi là **giá trị xác suất** hay **p - giá trị (p - value)** của phép kiểm định.

Phương pháp so sánh cặp đôi

Trong phụ lục này, chúng tôi chỉ để cập tới phép kiểm định được sử dụng trong luận án đó là *Phương pháp so sánh cặp đôi*. Phần này sẽ trình bày cách thực hiện phép kiểm định giả thuyết cho sự khác nhau giữa trung bình cặp đôi (paired means). Thủ tục kiểm định được sử dụng có tên là **matched-pairs t-test**, gồm 4 bước như sau:

1. Xác định các giả thuyết

   Các giả thuyết liên quan tới một biến d, biến này dựa trên sự khác nhau giữa các giá trị cặp đôi từ hai tập dữ liệu.

   $$d = x_1 - x_2$$

   với x_1 là giá trị của biến x trong tập dữ liệu thứ nhất, x_2 là giá trị của biến x trong tập dữ liệu thứ hai và có quan hệ cặp đôi với x_1.

   Bảng dưới đây chỉ ra ba tập giả thuyết không và đối thuyết; mỗi tập tạo nên một phát biểu về mối quan hệ giữa sự khác nhau thực sự trong giá trị tổng thể μ_d và giá trị giả thuyết D.

Tập	**Giả thuyết không**	**Đối thuyết**
1	$\mu_d = D$	$\mu_d \neq D$
2	$\mu_d \geq D$	$\mu_d < D$
3	$\mu_d \leq D$	$\mu_d > D$

2. Xây dựng kế hoạch phân tích

   Bước này mô tả cách sử dụng dữ liệu mẫu để chấp nhận hoặc từ chối giả thuyết không. Trong bước này cần xác định các thành phần sau:

   - Mức ý nghĩa (Thông thường, các nhà nghiên cứu chọn mức ý nghĩa là 0.01, 0.05, hoặc không 0.10.)
   - Phương pháp kiểm định (Sử dụng matched-pairs t-test để xác định xem sự khác nhau giữa các trung bình mẫu cho dữ liệu cặp đôi có thực sự khác sự khác nhau giả thuyết giữa các trung bình tổng thể.)

3. Phân tích dữ liệu mẫu

   Sử dụng dữ liệu mẫu để tìm độ lệch chuẩn, lỗi chuẩn, độ tự do (degrees of freedom), kiểm định thống kê, và giá trị P gắn với kiểm định thống kê.

- Độ lệch chuẩn: Tính độ lệch chuẩn (s_d) của sự khác nhau được tính từ n cặp đôi

$$s_d = sqrt[(\Sigma(d_i - \overline{d})^2/(n-1)]$$

trong đó d_i là sự khác nhau cho cặp i, $\overline{d}$ là sự khác nhau trung bình trên mẫu, và n là số cặp đôi.

- Lỗi chuẩn: Tính lỗi chuẩn (SE) của phấn phối lấy mẫu của $\overline{d}$

$$SE = s_d * sqrt(1/n) * (1 - n/N) * [N/(N-1)]$$

trong đó s_d là độ lệch chuẩn của sự khác nhau trên mẫu, N là kích thước tổng thể, n là kích thước mẫu. Khi kích thước tổng thể lớn hơn nhiều (lớn hơn ít nhất 10 lần) so với kích thước mẫu thì lỗi chuẩn có thể được tính xấp xỉ như sau:

$$SE = s_d/sqrt(n)$$

- Độ tự do $DF = n - 1$
- Kiểm định thống kê: Kiểm định thống kê là *điểm t* (t-score) được định nghĩa bởi:

$$t = [(\overline{x_1} - \overline{x_2}) - D]/SE = (\overline{d} - D)/SE$$

trong đó $\overline{x_1}$ là trung bình của mẫu 1, $\overline{x_2}$ là trung bình của mẫu 2, $\overline{d}$ là sự khác nhau trung bình giữa các giá trị cặp đôi trong tập mẫu, D là sự khác nhau giả thuyết giữa các trung bình tổng thể, SE là lỗi chuẩn.

- Giá trị P: Từ *điểm t* tính được ở trên và *Độ tự do DF* sẽ có được xác suất P tương ứng.

4. Giải thích kết quả

So sánh giá trị P với mức ý nghĩa và từ chối giả thuyết không khi giá trị P nhỏ hơn mức ý nghĩa.