FACIAL ANIMATION TECHNIQUES FOR VIETNAMESE PRONUNCIATION

Do Thi Chi[1*], Le Van Thuy[2], Le Son Thai[1], Ma Van Thu[1]
[1]TNU - University of Information and Communication Technology
[2]TNU - School of Foreign Language

ARTICLE INFO

Received: 14/02/2023
Revised: 07/4/2023
Published: 13/4/2023

KEYWORDS

Facial animation
Vietnamese pronunciation
Virtual reality
Audio-driven
3D animation

ABSTRACT

The automatic facial animation system according to the voice contributes to reducing the time and effort for character animators in the construction of cartoons, simulation graphics systems and virtual reality. Based on previous studies on existing animation techniques and specific features in Vietnamese pronunciation, we build a Vietnamese animation translator as the basis for automatic facial expression. This technique allows to convert Vietnamese text into virtual facial speech movements that synchronize with audio in real time. The animation image of the virtual character is evaluated based on the perception of real users and experts in the field of animation with the criteria of realism, naturalness and smoothness for good results. Since then, the animation technique proposed in the article allows to replace or partially support the work of facial animation of three-dimensional characters with Vietnamese. At the same time, this study helps to improve the communication method between humans and computers.

KỸ THUẬT DIỄN HOẠT KHUÔN MẶT THEO PHÁT ÂM TIẾNG VIỆT

Đỗ Thị Chi[1*], Lê Văn Thủy[2], Lê Sơn Thái[1], Mã Văn Thu[1]
[1]Trường Đại học Công nghệ Thông tin và Truyền thông – ĐH Thái Nguyên
[2]Trường Ngoại ngữ - ĐH Thái Nguyên

THÔNG TIN BÀI BÁO

Ngày nhận bài: 14/02/2023
Ngày hoàn thiện: 07/4/2023
Ngày đăng: 13/4/2023

TỪ KHÓA

Diễn hoạt khuôn mặt
Phát âm tiếng Việt
Thực tại ảo
Điều khiển theo âm thanh
Diễn hoạt ba chiều

TÓM TẮT

Hệ thống diễn hoạt khuôn mặt tự động theo tiếng nói góp phần giảm thời gian và công sức cho các nhà diễn hoạt nhân vật trong xây dựng phim hoạt hình, hệ thống đồ họa mô phỏng và thực tại ảo. Dựa trên những nghiên cứu về các kỹ thuật diễn hoạt đã có và các đặc trưng riêng biệt trong phát âm tiếng Việt, chúng tôi xây dựng bộ dịch hoạt Việt là cơ sở cho diễn hoạt tự động khuôn mặt. Kỹ thuật này cho phép chuyển đổi văn bản tiếng Việt thành các chuyển động nói trên khuôn mặt ảo đồng bộ với âm thanh trong thời gian thực. Hình ảnh diễn hoạt của nhân vật ảo được đánh giá dựa trên cảm nhận của người dùng thực và các chuyên gia trong lĩnh vực diễn hoạt với các tiêu chí về độ chân thực, tự nhiên và độ mịn cho kết quả tốt. Từ đó, kỹ thuật diễn hoạt đã đề xuất trong nội dung bài báo cho phép thay thế hoặc hỗ trợ một phần trong công việc diễn hoạt trên khuôn mặt của nhân vật ba chiều với tiếng Việt. Đồng thời, nghiên cứu này giúp hoàn thiện hơn phương thức giao tiếp giữa người và máy tính.

DOI: https://doi.org/10.34238/tnu-jst.7332

[*] Corresponding author. *Email: dtchi@ictu.edu.vn*

1. Giới thiệu

Diễn hoạt khuôn mặt là một phần trong quy trình xây dựng các bộ phim hoạt hình hay các hệ thống giao tiếp giữa con người và máy tính. Trong các bộ phim hoạt hình, các nhân vật nói, trao đổi và đọc các lời thoại thông qua âm thanh và được đồng bộ với sự thay đổi của các bộ phận trên khuôn mặt. Trong các hệ thống thông minh và thực tại ảo, máy tính có khả năng nói chuyện và tương tác trực tiếp với con người qua nhân vật ảo ba chiều có khả năng thể hiện các chuyển động trên khuôn mặt đồng bộ với âm thanh phát ra. Các hệ thống diễn hoạt này phải đảm bảo sự đồng bộ giữa âm thanh và hình ảnh sinh ra, nhiều hệ thống đòi hỏi phải hoàn thiện quá trình tính toán và kết xuất hình ảnh trong thời gian thực. Việc giải quyết tốt bài toán diễn hoạt khuôn mặt theo âm thanh trong thời gian thực còn có nhiều ý nghĩa khi xây dựng các hệ thống giao tiếp trực tiếp giữa người và máy tính như giáo viên ảo, phát thanh viên ảo, trợ lý ảo v.v..

Có nhiều nghiên cứu khác nhau [1] cho việc diễn hoạt tự động khuôn mặt theo âm thanh với hai hướng tiếp cận chính. Hướng tiếp cận thứ nhất [2], [3] là diễn hoạt theo đối tượng người thực, tại đó hình ảnh con người khi nói được máy tính thu lại dưới dạng hình ảnh trực tiếp [4], [5] hoặc video và là dữ liệu đầu vào cho các quá trình trích lọc thông tin hay học máy [6], [7] sau đó. Hướng tiếp cận thứ hai là dựa trên các đặc điểm của từng ngôn ngữ, các chuyên gia về diễn hoạt xây dựng các bộ quy tắc điều khiển từ đó diễn hoạt khuôn mặt được tạo ra từ các bộ quy tắc này. Cả hai hướng tiếp cận này đều tồn tại các ưu điểm và nhược điểm khác nhau.

Với hướng tiếp cận thứ nhất, một tập dữ liệu mẫu đầu vào [6], [7] được xây dựng dựa trên người thực. Tại đó, người làm mẫu đọc và thể hiện các lời thoại chuẩn bị trước để tạo ra các cử động trên khuôn mặt và được ghi lại thông qua camera hoặc các thiết bị cảm biến khác nhau [5]. Ưu điểm của hướng tiếp cận này là việc sử dụng lại của tập dữ liệu sau khi xây dựng, khi đó các nhà nghiên cứu thay đổi các kỹ thuật trích rút thông tin hoặc mô hình học máy sẽ cho những kết quả đầu ra khác nhau. Nhược điểm của hướng tiếp cận này là dữ liệu đầu vào khó trích lọc chính xác thông tin, đồng thời các video hai chiều hay cảm biến chưa ghi lại hết các chuyển động trên khuôn mặt trong không gian thực. Bên cạnh đó việc gán nhãn chính xác dữ liệu theo thời gian cũng là một vấn đề từ đó gây khó khăn khi xây dựng một bộ dữ liệu học với kích thước lớn và thời gian dài. Bên cạnh đó, với các ngôn ngữ khác nhau đòi hỏi các bộ dữ liệu đầu vào khác nhau và để diễn hoạt khuôn mặt phù hợp cần chuẩn bị tốt bộ dữ liệu đầu vào.

Với hướng tiếp cận thứ hai [8] – [10], các đặc trưng của ngôn ngữ được các chuyên gia diễn hoạt phân tích và đưa ra các bộ quy tắc diễn hoạt khi nói. Hệ thống diễn hoạt khuôn mặt tự động tuân theo các quy tắc diễn hoạt này và các tham số khác như độ lớn của âm thanh, tình cảm của nhân vật v.v.. để kết xuất ra các trạng thái khác nhau của khuôn mặt theo dòng thời gian. Ưu điểm của hướng tiếp cận này là tạo ra được một bộ quy tắc chung khi diễn hoạt và không phụ thuộc nhiều vào bộ dữ liệu đầu vào. Nhược điểm chính của hướng tiếp cận này là độ chính xác phụ thuộc vào chuyên gia khi đưa ra bộ quy tắc khi diễn hoạt. Khi đó để nâng cao kết quả khi diễn hoạt cần phải chỉnh sửa lại bộ quy tắc diễn hoạt và các kỹ thuật xử lý liên quan.

Hiện nay, các kỹ thuật diễn hoạt khuôn mặt được các nhà nghiên cứu quan tâm và đã có những đề xuất cho từng ngôn ngữ khác nhau [10], [11]. Các phần mềm thiết kế phổ dụng như Maya, FaceFX, 3ds Max có các phần hỗ trợ tạo hoạt ảnh môi bằng nhiều ngôn ngữ nước ngoài như tiếng Anh, tiếng Hàn, tiếng Pháp… nhưng chưa có các hỗ trợ diễn hoạt giọng nói tiếng Việt cho nhân vật 3D. Điều này thúc đẩy các nghiên cứu về diễn hoạt khuôn mặt cho tiếng nói Việt.

Trong phạm vi bài báo, chúng tôi tìm hiểu các nghiên cứu về lĩnh vực diễn hoạt khuôn mặt nói chung và đưa ra một kỹ thuật diễn hoạt với tiếng nói Việt có khả năng thực hiện trong thời gian thực. Tiếng nói Việt mang những đặc điểm riêng mà không ngôn ngữ nào có trên thế giới. Theo đó, chúng tôi đề xuất kỹ thuật diễn hoạt dựa trên việc phân tích phát âm trong tiếng nói Việt và từ đó xây dựng một hệ thống tự động diễn hoạt theo âm thanh khi nói của nhân vật ảo. Cùng với sự phát triển của việc tổng hợp tiếng nói nhân tạo [12] và các công nghệ trí tuệ nhân tạo [13]

cho phép máy tính có thể trả lời và giao tiếp tự động. Nghiên cứu này góp phần vào quá trình hoàn thiện các hệ thống tương tác trực tiếp giữa người và máy tính.

2. Một số kỹ thuật diễn hoạt khuôn mặt

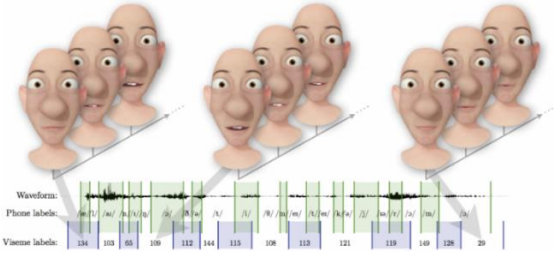
Có nhiều các kỹ thuật diễn hoạt khuôn mặt khác nhau [1] đã được các nhà nghiên cứu tìm hiểu và đề xuất. Trong đó, với hướng tiếp cận thu thập hình ảnh người thực khi nói làm đầu vào cho quá trình diễn hoạt được nhiều nhà phát triển sử dụng.

Choong Seng Chan và các cộng sự [2] đã sử dụng kỹ thuật PCA và EMPCA để trích lọc các thông tin trên khuôn mặt từ video người nói thực. Quá trình xử lý đưa ra tập các điểm đánh dấu trên khuôn mặt và sự thay đổi vị trí của các điểm này tạo ra sự thay đổi trạng thái trên khuôn mặt ảo. Thay vì sử dụng 50 điểm như một số tiếp cận khác Choong Seng Chan và các cộng sự sử dụng 103 điểm cho quá trình diễn hoạt và kết quả được đánh giá là tự nhiên và giống thật hơn. Đi cùng với các kỹ thuật xử lý ảnh và nhận dạng hình ảnh, hướng tiếp cận dựa trên xử lý hình ảnh thu nhận từ người thực tiếp tục được các nhà nghiên cứu sử dụng. Yanxiang Zhang và Yan Ling [3] đã tạo ra các ảnh đại diện động diễn hoạt các biểu cảm khác nhau của khuôn mặt từ video của người sử dụng. Changwei Luo và các cộng sự [4] đưa ra một quy trình cho việc diễn hoạt mặt trong thời gian thực từ việc xử lý hình ảnh trực tiếp từ máy quay. Một số nhà nghiên cứu khác sử dụng Kinect [5] khi thu nhận hình ảnh của khuôn mặt. Với các thông tin về chiều sâu trong không gian ba chiều, Kinect giúp quá trình phân tích các chuyển động trên khuôn mặt nhanh và hiệu quả hơn. Theo đó, quá trình thu nhận và diễn hoạt khuôn mặt đảm bảo trong thời gian thực.

Bên cạnh đó, với sự phát triển của trí tuệ nhân tạo. Học máy đã được sử dụng nhiều trong diễn hoạt khuôn mặt. Karras và cộng sự [6] đã sử dụng kỹ thuật học máy để diễn hoạt khuôn mặt 3D theo âm thanh trong thời gian thực. Theo đó, máy học là một mạng neural đã được xây dựng để học tập các biểu cảm và chuyển động từ người thực kết hợp với âm thanh. Nó cho phép diễn hoạt khuôn mặt từ tập dữ liệu học tập là video với độ trễ thấp khi vận hành. Các hệ thống học máy, học sâu khi diễn hoạt khuôn mặt phụ thuộc vào dữ liệu đầu vào dùng để huấn luyện và kiến trúc của mô hình. Khi dữ liệu huấn luyện không đủ nhiều hoặc không phù hợp với ngôn ngữ được sử dụng mang tới những hạn chế trong quá trình diễn hoạt. Nhận thấy điều đó, Daniel Cudeiro và các cộng sự [7] đã giới thiệu bộ dữ liệu khuôn mặt 4D với khoảng 29 phút quét 4D được chụp ở tốc độ 60 khung hình/giây và âm thanh được đồng bộ hóa từ 12 người đọc. Mô hình máy học sau đó cho phép lấy tín hiệu lời nói làm đầu vào và tạo hoạt ảnh cho khuôn mặt người lớn.

Ở một hướng tiếp cận khác [8] – [10], các diễn hoạt trên khuôn mặt ảo được tạo ra bởi các quy tắc phát âm kết hợp với mô hình sinh hình ảnh trung gian. Với cách tiếp cận này cho phép nhân vật ảo tạo ra các chuyển động trên khuôn mặt dựa trên một đoạn âm thanh đầu vào được gán nhãn. Tại đó, các từ trong câu được phân tích thành các nhóm âm thanh tương ứng với các trạng thái khác nhau của khuôn mặt được gọi là viseme.

Taylor và các cộng sự [8] đã định nghĩa các viseme mô tả các chuyển động theo lời nói của người phát âm. Trong đó, các âm vị phân thành các nhóm theo hình dạng miệng tĩnh và được đại diện bởi một viseme (ví dụ: /p, b, m/, và /f, v/). Những biểu diễn giống nhau khi phát âm được nhóm lại và, từ đó cung cấp một tập hợp hữu hạn các chuyển động mô tả trạng thái khi nói. Hình ảnh động thu được khi nói bằng cách nối các đơn vị hình ảnh động lại với nhau.



Hình 1. *Diễn hoạt khuôn mặt sử dụng cụm các âm vị [8]*

Hình 1 mô tả quy trình mà Taylor và các cộng sự đã sử dụng khi diễn hoạt khuôn mặt. Sóng âm thanh được gán nhãn tương ứng với các viseme để tạo ra diễn hoạt theo dòng thời gian.

P. Edwards và các cộng sự [9] sử dụng quan hệ của môi và hàm đã đề xuất kỹ thuật JALI cho quá trình diễn hoạt khuôn mặt theo âm thanh đầu vào. Tại đó, các thay đổi trong diễn hoạt không chỉ phụ thuộc vào việc nhóm các âm vị khác nhau và hình thái của chúng mà còn phụ thuộc vào các tham số điều khiển của môi và hàm được nhóm tác giả gọi là tham số JA và Li.

Hướng tiếp cận dựa trên tập các quy tắc phát âm cho hình ảnh diễn hoạt có kết quả phụ thuộc vào cách nhóm các âm vị và hệ thống điều khiển sinh hình ảnh nối tiếp giữa các âm vị này. Bên cạnh đó, với mỗi loại ngôn ngữ khác nhau thì cách phát âm cũng có sự khác biệt tương đối. Điều này dẫn tới các nghiên cứu riêng cho quá trình phát âm của mỗi ngôn ngữ khác nhau [10], [11].

Đối với ngôn ngữ tiếng Việt, có một số đề xuất cho diễn hoạt khuôn mặt. Tuy nhiên, các nhà nghiên cứu thường quan tâm tới biểu cảm khuôn mặt nhiều hơn để thể hiện cảm xúc khi nói và các quy tắc khi phát âm còn đơn giản. T. D. Ngo và các cộng sự [14] đã đề xuất hệ thống diễn hoạt cho ngôn ngữ Việt với các viseme được phân chia dựa trên các nguyên âm và phụ âm kết hợp với nghiên cứu về sinh tiếng nói và biểu diễn biểu cảm để tạo các diễn hoạt cảm xúc. Hình 2 bên dưới là kết quả đạt được khá khi áp dụng bộ quy tắc diễn hoạt dựa trên nguyên âm và phụ âm kết hợp với thể hiện cảm xúc khi nói cho ngôn ngữ tiếng Việt.

Hình 2. *Diễn hoạt tiếng nói Việt có thể hiện cảm xúc [14]*

Trong thời gian gần đây, với sự phát triển không ngừng của các hệ thống nhận dạng, tổng hợp tiếng nói [12] và trí tuệ nhân tạo [13] vấn đề diễn hoạt khuôn mặt đang được các nhà nghiên cứu quan tâm và có nhiều đề xuất khác nhau. Điều này tạo ra một hệ thống đồng bộ trong giao tiếp giữa người và máy tính. Trong đó, quá trình diễn hoạt trên khuôn mặt đòi hỏi đồng bộ với âm thanh và thường diễn ra trong thời gian thực. Đồng thời, cần có những nghiên cứu đặc trưng riêng cho từng ngôn ngữ để tạo ra diễn hoạt chính xác theo từng dạng tiếng nói.

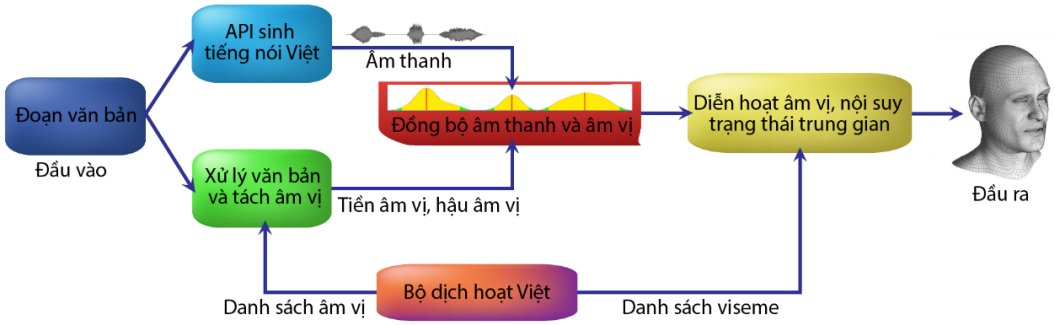
3. Kỹ thuật diễn hoạt khuôn mặt theo phát âm tiếng Việt và đánh giá kết quả

Quá trình diễn hoạt khuôn mặt thường được các nhà thiết kế và chuyên gia diễn hoạt thực hiện. Với đầu vào là một đoạn âm thanh hội thoại, các nhân vật ảo sẽ có những thể hiện tương ứng với ngữ nghĩa của từ và tình cảm của nhân vật. Tuy nhiên, quá trình diễn hoạt thủ công này tốn nhiều thời gian, công sức và phụ thuộc vào kỹ năng của người thực hiện. Bên cạnh đó, với sự phát triển của các kỹ thuật tạo tiếng nói tự động từ văn bản [12] cũng như các kỹ thuật về xử lý hình ảnh dẫn tới các kỹ thuật diễn hoạt trong thời gian thực [1] và tự động theo âm thành ngày càng được quan tâm. Đối với diễn hoạt ngôn ngữ tiếng Việt hướng tiếp cận dựa trên xử lý hình ảnh trực tiếp và mạng neural gặp hạn chế khi tập dữ liệu mẫu được gán nhãn ít ảnh hưởng tới kết quả hệ thống diễn hoạt tự động. Đồng thời với những đặc trưng riêng của ngôn ngữ Việt, hướng tiếp cận dựa trên tập quy tắc diễn hoạt có độ phù hợp cao hơn. Bằng việc sử dụng các API chuyển đổi văn bản thành giọng nói và phân tích âm vị chúng tôi đề xuất kỹ thuật diễn hoạt khuôn mặt cho ngôn ngữ tiếng Việt theo bộ dịch hoạt Việt theo quy trình tại hình 3 bên dưới.

Đầu vào của hệ thống diễn hoạt là một đoạn văn bản do người sử dụng lựa chọn bằng ngôn ngữ tiếng Việt. Tiếp đó là hai luồng xử lý song song cho việc sinh tiếng nói tự động và xử lý văn bản. Đối với quá trình sinh tiếng nói Việt từ văn bản, chúng tôi sử dụng các API được cung cấp sẵn cho phép chuyển đổi văn bản thành giọng nói.

Quá trình xử lý văn bản và tách âm vị dựa trên đặc trưng của ngôn ngữ tiếng Việt là ngôn ngữ đơn âm. Mỗi âm phát ra thường có thể chia thành hai hình thái là trước và sau khi phát âm gọi là

tiền âm vị và hậu âm vị. Trong đó, tiền âm vị là trạng thái xuất phát, hậu âm vị là trạng thái kết thúc khi nói một âm. Quá trình tách âm vị trong câu văn bản dựa trên việc rà soát từ cuối xâu ký tự lên đầu dựa trên từ điển hậu âm vị do bộ dịch hoạt cung cấp. Hậu âm vị xác định được là âm dài nhất có trong từ điển. Trên thực tế, có những từ đơn chỉ có hậu âm vị mà không có tiền âm vị và ngược lại. Hình 4 minh họa ví dụ về quy trình xử lý văn bản gồm loại bỏ dấu và tách âm vị. Trong đó, tiền âm vị và hậu âm vị phân biệt bởi dấu gạch dưới.



Hình 3. *Kỹ thuật diễn hoạt khuôn mặt theo phát âm tiếng Việt*

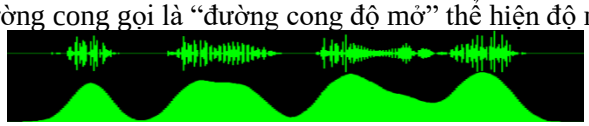
Hình 4. *Ví dụ về xử lý câu đầu vào*

Bộ dịch hoạt cung cấp danh sách tiền âm vị và hậu âm vị cùng các viseme tương ứng. Để xây dựng bộ dịch hoạt các hình dáng khuôn mặt được xác định dựa trên các chuyên gia về diễn hoạt và các nghiên cứu về phát âm tiếng Việt. Chúng tôi tìm hiểu các tài liệu về giảng dạy tiếng Việt và xây dựng bộ dịch hoạt của tiếng Việt gồm: 12 nguyên âm đơn, 139 nguyên âm ghép, 17 phụ âm đơn và 11 phụ âm ghép. Tuy nhiên, phụ âm và thanh điệu là âm thanh của lời nói, được phát âm rõ ràng với sự đóng hoàn toàn hay một phần của thanh quản. Do đó, trong 28 phụ âm, chúng tôi diễn hoạt 4 phụ âm b/m/p/ph vì chỉ có 4 phụ âm này có sự tác động nhiều của môi. Các thanh điệu có ảnh hưởng tới hình ảnh khi diễn hoạt, nhưng sự ảnh hưởng này là không lớn. Với mục tiêu tối ưu bộ dịch hoạt, các thanh điệu được sử dụng trong tiếng Việt được loại bỏ khi diễn hoạt. Đây là nguyên nhân phần dấu bị loại bỏ trong quá trình xử lý văn bản. Đồng thời, khi cố gắng xây dựng bộ dịch hoạt tiếng Việt với đầy đủ các dấu sẽ phát sinh sự bùng nổ tổ hợp. Tựu chung lại, trong bài báo chúng tôi liệt kê 154 âm vị. Với những âm vị có hình dáng chuyển động của môi giống nhau sẽ được gộp thành 1 nhóm và xây dựng viseme tương ứng. Hình 5 là một số viseme của bộ dịch hoạt tiếng Việt.

Hình 5. *Một số viseme trong bộ dịch hoạt Việt*

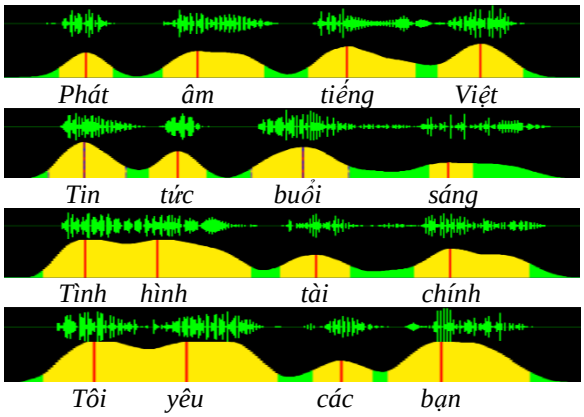
Quá trình đồng bộ giữa âm thanh và âm vị là việc thiết đặt vị trí của tiền âm vị và hậu âm vị trên trục thời gian của âm thanh. Quá trình diễn hoạt một câu là kết hợp của việc diễn hoạt từng âm trong câu khi nói. Trong đó, khoảng cách giữa mỗi âm là trạng thái tự nhiên của nhân vật 3D khi chưa nói. Quá trình diễn hoạt một âm là kết hợp của việc chuyển đổi trạng thái từ tiền âm vị tới hậu âm vị.

Trong quá trình đồng bộ cần phân tích đặc trưng âm thanh để xác định độ mở tương ứng của miệng và vị trí của các âm vị. Do sử dụng âm thanh sinh ra từ API tạo tiếng nói tự động nên âm thanh không có tạp âm và các từ được phân tách rõ ràng tạo thuận lợi khi phân tích. Đầu tiên, quá trình khép mở của miệng tương đồng với sự thay đổi biên độ dao động của âm thanh. Thường thì độ mở tỉ lệ thuận với độ lớn của âm thanh khi nói. Chúng tôi trích chọn đặc trưng dao động này và làm mịn thành một đường cong gọi là "đường cong độ mở" thể hiện độ mở của miệng.



Hình 6. *Đặc trưng độ mở của miệng khi nói "Phát âm tiếng Việt"*

Hình 6 biểu diễn đường cong đặc trưng cho độ mở của miệng khi nói câu "Phát âm tiếng Việt". Dựa trên quá trình lấy mẫu, trích chọn đặc trưng và làm mịn, đường cong độ mở thu được thay đổi theo thời gian và được sử dụng để điều khiển độ mở của miệng. Bên cạnh đó, dựa theo sự thay đổi trên đường cong độ mở chúng tôi xác định vị trí của các âm trên dòng thời gian khi nói.



Hình 7. *Vị trí âm vị trên đường cong độ mở của miệng với một số câu khác nhau*

Hình 7 mô tả vị trí của các âm vị được xác định trên đường cong độ mở. Vạch đỏ là vị trí đỉnh của một đoạn cong tương ứng với mỗi âm được coi là trọng tâm của từng âm. Phần màu vàng là vùng có tiếng nói tương ứng trong câu, qua đó hỗ trợ xác định được vị trí bắt đầu và kết thúc khi nói của một âm. Trong một câu có N âm tương ứng với N vị trí là đỉnh của một đoạn cong. Các đỉnh này là các vị trí cao nhất của các đoạn cong. Giải pháp lựa chọn đỉnh này cho kết quả tương đối tốt trong điều kiện âm thanh đầu vào không có tạp âm và tốc độ nói ở mức bình thường. Kỹ thuật xử lý âm thanh này phù hợp với việc sử dụng API tạo giọng nói tự động. Trong trường hợp quá trình xử lý âm thanh không đáp ứng được yêu cầu, chúng tôi sử dụng các thông tin do API cung cấp để có được thông tin tương đối về vị trí trọng tâm của mỗi âm. Vị trí tiền âm vị được xác định là phần có tiếng nói ở phía trước mỗi trọng tâm của âm và ngược lại hậu âm vị thuộc về phần có tiếng nói sau trọng tâm. Trong trường hợp phần tiếng nói liên tiếp nhau vị trí chính giữa được lựa chọn là vị trí phân tách giữa 2 âm.

Quá trình diễn hoạt âm vị và nội suy trạng thái trung gian là sự chuyển đổi từ tiền âm vị tới hậu âm vị trong khoảng thời gian mà âm được xác định trên đường cong độ mở. Quá trình này là sự đồng bộ từ độ mở của miệng tới việc chuyển đổi trạng thái khuôn mặt từ tiền âm vị tới hậu âm vị. Khi bắt đầu một âm, trạng thái khuôn mặt chuyển đổi từ trạng thái bình thường sang tiền âm vị tới hậu âm vị và quay trở lại trạng thái bình thường. Với những âm không có tiền âm vị hoặc hậu âm vị, quá trình này đơn giản hơn khi chuyển từ trạng thái bình thường sang âm vị xác định được nằm ở trọng tâm của âm và về trạng thái bình thường. Khi phần tiếng nói liên tiếp nhau trạng thái bình thường được bỏ qua và chuyển từ hậu âm vị của âm trước sang tiền âm vị của âm sau. Một câu là việc lặp lại quá trình diễn hoạt với từng âm theo thứ tự.

Quá trình chuyển đổi giữa các âm vị và các trạng thái sử dụng nội suy để tính toán trạng thái giữa theo thời gian. Với mỗi âm, tiền âm vị được đặt ở trung điểm của vị trí bắt đầu và trọng tâm của nó trên đường cong độ mở. Hậu âm vị được đặt ở trung điểm vị trí kết thúc âm và trọng tâm trên đường cong độ mở. Các âm vị được sắp xếp trên dòng thời gian, trạng thái diễn hoạt của nhân vật tại thời gian t là nội suy của 2 âm vị gần nhất, trước và sau thời gian t. Giả sử, thứ tự các âm vị là p_0, p_1,…, p_n diễn ra tại các thời điểm t_0, t_1,…, t_n. Ta xác định được vị trí của t trên dòng thời gian với điều kiện: $t_i \le t \le t_{i+1}$. Trạng thái p_t mô tả hình dáng miệng tại thời gian t được xác định dựa trên hàm nội suy tuyến tính (1).

$$p_t = \frac{t - t_i}{t_{i+1} - t_i} p_{i+1} + \frac{t_{i+1} - t}{t_{i+1} - t_i} p_i \quad (1)$$

Để đảm bảo t luôn nằm giữa hai trạng thái khi nói, trạng thái nhân vật khép miệng được đưa vào đầu và cuối dòng âm thanh. Quá trình làm mịn dựa trên nội suy giữa hai trạng thái gần đảm bảo hoạt ảnh sinh ra liên tục giúp hình ảnh sinh ra đạt độ mịn trong thời gian thực. Tổng hợp các quá trình xử lý chúng ta thu được hình ảnh diễn hoạt nói tiếng Việt của nhân vật 3D với đầu vào là một đoạn văn bản tiếng Việt và đầu ra là hình ảnh diễn hoạt của nhân vật ảo tương ứng.

Quá trình nói của con người là sự tổng hợp của nhiều bộ phận. Do đó, để nâng cao tính chân thực khi diễn hoạt nói cho nhân vật ảo chúng tôi tiến hành điều khiển với một số bộ phận: lưỡi, xương hàm dưới, thanh quản, lồng ngực khi nói. Điều này nâng cao hình ảnh diễn hoạt một cách tổng thể và tạo chiều sâu giải phẫu cho nhân vật ảo. Hình 8 là một số bộ phận trên cơ thể ngoài khuôn mặt được chúng tôi sử dụng khi tham gia vào quá trình diễn hoạt.

Hình 8. *Sử dụng một số bộ phận phát âm trên cơ thể tăng độ chân thực khi diễn hoạt nói*

Để đánh giá các kết quả của kỹ thuật diễn hoạt, các nhà nghiên cứu thường sử dụng cảm nhận của chính con người. C. Luo và các cộng sự [4] sử dụng tiêu chí về độ tự nhiên, độ chân thực và độ mịn để đánh giá kết quả diễn hoạt. T. Karras và các cộng sự [6] sử dụng 20 người tình nguyện, C. Weerathunga và các cộng sự [10] sử dụng 30 sinh viên đại học để đánh giá kết quả diễn hoạt của nhân vật ảo. Trong nghiên cứu này, chúng tôi sử dụng hai nhóm tình nguyện. Nhóm đầu tiên là những người dùng bình thường để đánh giá kết quả cảm nhận của người dùng. Nhóm thứ hai gồm các chuyên gia có kiến thức về diễn hoạt để đánh giá kết quả cũng như khả năng hỗ trợ của kỹ thuật diễn hoạt tự động được đề xuất đối với công việc của họ. 30 đoạn video là hình ảnh diễn hoạt của nhân vật ảo khi đọc đoạn giới thiệu của các bài tin tức trên báo điện tử Việt Nam được sử dụng làm dữ liệu đánh giá.

Bảng 1. *Đánh giá 30 đoạn video hình ảnh diễn hoạt nhân vật ảo khi đọc tin tức tiếng Việt*

Nhóm	Thành phần	Số lượng	Tiêu chí đánh giá			
			Chân thực	Tự nhiên	Độ mịn	Hỗ trợ, thay thế
1	Người dùng thường	40	80,25%	72%	82,25%	-
2	Chuyên gia	20	78,5%	65,5%	81,25%	67,5%

Bảng 1 cho thấy, đối với người dùng bình thường, các tiêu chí đánh giá thường có giá trị cao hơn so với đánh giá của người có chuyên môn trong diễn hoạt. Trong đó, độ mịn được đánh giá cao khi sự chuyển đổi các khung hình và đồng bộ với âm thanh đạt được các thông số kỹ thuật tốt. Độ tự nhiên của hình ảnh sinh ra được đánh giá thấp hơn độ chân thực do nhân vật ảo do chưa thể hiện được cảm xúc khi nói. Khả năng hỗ trợ thay thế được sử dụng cho nhóm có chuyên

môn trong diễn hoạt để đánh giá khả năng giúp ích của diễn hoạt tiếng Việt tự động tới công việc của họ. Trung bình nhóm chuyên gia đánh giá khả năng giúp ích khoảng 67,5% so với khi họ thực hiện diễn hoạt thủ công.

4. Kết luận

Trong nội dung bài báo, chúng tôi đã trình bày kỹ thuật diễn hoạt khuôn mặt được các nhà nghiên cứu đề xuất và sử dụng thời gian gần đây. Tiếp đó, dựa trên các đặc trưng của ngôn ngữ tiếng Việt nhóm tác giả đưa ra kỹ thuật diễn hoạt với đầu vào là đoạn văn bản và đầu ra là hình ảnh diễn hoạt trên khuôn mặt được đồng bộ theo âm thanh. Chúng tôi đề xuất kỹ thuật sử dụng tiền âm vị và hậu âm vị được chia thành các nhóm với cùng viseme làm cơ sở cho quá trình diễn hoạt. Đề xuất này dựa theo các nghiên cứu về đặc trưng của tiếng nói Việt từ đó ứng dụng cho hệ thống diễn hoạt tự động trong thời gian thực trên bộ dịch hoạt Việt được chúng tôi xây dựng.

Các kết quả là hình ảnh và video diễn hoạt được đánh giá dựa trên cảm nhận của người dùng thực và các chuyên gia trong lĩnh vực diễn hoạt. Đối với người dùng bình thường ghi nhận mức đánh giá từ 72% tới 82,25% đối với các chỉ số về độ tự nhiên, chân thực và độ mịn của hình ảnh. Đối với các chuyên gia được đào tạo trong lĩnh vực diễn hoạt cho thấy hệ thống có khả năng đảm bảo 67,5% công việc của họ khi diễn hoạt các chuyển động trên khuôn mặt khi nói.

Dựa trên các kết quả đạt được, kỹ thuật diễn hoạt khuôn mặt tự động theo tiếng nói Việt cho phép hoàn thiện các hệ thống giao tiếp tự động giữa người và máy tính bằng tiếng Việt. Từ đó, hỗ trợ xây dựng các hệ thống thay thế con người trong trò chuyện, giao tiếp, giảng dạy v.v.. Tuy nhiên, cần hoàn thiện hơn các kỹ thuật sử dụng cho diễn hoạt khuôn mặt cho tiếng nói Việt để đảm bảo độ chính xác cao và tăng khả năng biểu cảm giống con người khi giao tiếp. Điều đó đòi hỏi cần có những nghiên cứu để diễn hoạt tự nhiên và thể hiện tính cách cũng như tiệm cận tới cách diễn đạt của con người khi nói.

Lời cảm ơn

Bài báo là sản phẩm của đề tài mã số T2022-07-18 do Trường Đại học Công nghệ thông tin và Truyền thông cấp kinh phí.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] J. Hwang and K. Park, "Audio-driven Facial Animation: A Survey," in *The 13th International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju Island, Korea, 2022, pp. 614-617.

[2] C. S. Chan and F. S. Tsai, "Computer Animation of Facial Emotions," in *International Conference on Cyberworlds*, Singapore, 2010, pp. 425-429.

[3] Y. Zhang and Y. Ling, "Interactive Narrative Facial Expression Animation Generation by Intuitive Curve Drawing," in *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, Lisbon, Portugal, 2021, pp. 406-409.

[4] C. Luo, J. Yu, C. Jiang, R. Li, and Z. Wang, "Real-time control of 3D facial animation," in *IEEE International Conference on Multimedia and Expo (ICME)*, Chengdu, China, 2014, pp. 1-6.

[5] C. Chen, Y. Zhang, P. Xu, S. Lan, S. Li, and Y. Zhang, "Real-time 3D facial expression control based on performance," in *12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Zhangjiajie, China, 2015, pp. 1324-1328.

[6] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Transactions on Graphics*, vol. 36, no.4, pp. 1–12, 2017.

[7] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, Learning, and Synthesis of 3D Speaking Styles," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR*), Long Beach, CA, USA, 2019, pp. 10093-10103.

[8] S. Taylor, M. Mahler, B.-J. Theobald, and I. Matthews, "Dynamic units of visual speech," *Proceedings of the 11th ACM SIGGRAPH/ Eurographics conference on Computer Animation*, 2012, pp. 275-284.

[9] P. Edwards, C. Landreth, E. Fiume, and K. Singh, "JALI: An animatorcentric viseme model for expressive lip synchronization," *ACM Transactions on Graphics*, vol. 35, no. 4, pp. 1–11, 2016.

[10] C. Weerathunga, R. Weerasinghe, and D. Sandaruwan, "Lip Synchronization Modeling for Sinhala Speech," in *20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka, 2020, pp. 208-213.

[11] R. Kato, Y. Kikuchi, V. Yem, and Y. Ikei, "CV-Mora Based Lip Sync Facial Animations for Japanese Speech," in *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, Christchurch, New Zealand, 2022, pp. 558-559.

[12] H.-T. Dang, T.-H.-Y. Vuong, and X.-H. Phan, "Non-Standard Vietnamese Word Detection and Normalization for Text–to–Speech," in *14th International Conference on Knowledge and Systems Engineering (KSE)*, Nha Trang, Vietnam, 2022, pp. 1-6.

[13] T. Kinouchi and N. Kitaoka, "A response generation method of chat-bot system using input formatting and reference resolution," in *9th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, Tokoname, Japan, 2022, pp. 1-6.

[14] T. D. Ngo and T. D. Bui, "A Vietnamese 3D taking face for embodied conversational agents," in *The 2015 IEEE RIVF International Conference on Computing & Communication Technologies - Research, Innovation, and Vision for Future (RIVF)*, Can Tho, Vietnam, 2015, pp. 94-99.