

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



TRẦN MINH TUẤN

**NGHIÊN CỨU TỔNG HỢP TIẾNG NÓI
VÀ ỨNG DỤNG ĐỌC BÁO BẰNG TIẾNG VIỆT TRÊN ĐIỆN
THOẠI ANDROID**

CHUYÊN NGÀNH : HỆ THỐNG THÔNG TIN

MÃ SỐ: 60.48.01.04

LUẬN VĂN THẠC SĨ KỸ THUẬT
(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS. LÊ HỮU LẬP

HÀ NỘI - 2016

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS.TS. Lê Hữu Lập

Phản biện 1:

Phản biện 2:

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ... giờ ngày tháng năm

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

Ngày nay, với sự phát triển như vũ bão của công nghệ thông tin, Internet cũng như các dịch vụ trực tuyến ngày càng có nhiều thông tin được tạo ra. Ta có thể truy cập các thông tin đó qua sách, báo, Internet và các phương tiện truyền thông. Cùng với đó là sự phát triển mạnh của các thiết bị di động Android. Ta có thể thu thập thông tin ở bất cứ nơi đâu thông qua thiết bị di động này. Hơn nữa, nhu cầu đọc, tìm hiểu và lưu trữ thông tin của con người ngày càng tăng lên. Tuy nhiên, với số lượng lớn thông tin như vậy thì ta không có đủ thời gian và sức lực để tiếp thu bằng phương pháp đọc thông thường. Giải pháp tổng hợp những thông tin dưới dạng văn bản này thành tiếng nói để cung cấp thêm một phương thức tiếp thu thông tin.

Tổng hợp tiếng nói là quá trình tạo ra tiếng nói nhân tạo của người trên máy tính từ văn bản. Đây là một đề tài có tính ứng dụng thực tiễn cao nên được nghiên cứu nhiều trên thế giới và Việt Nam từ rất sớm [2]. Tuy nhiên, chất lượng tiếng nói tổng hợp sao cho dễ nghe và tự nhiên vẫn là điều mà các công trình nghiên cứu đang hướng tới [6].

Vì vậy, Học viên xin chọn đề tài “ *Nghiên cứu tổng hợp tiếng nói và ứng dụng đọc báo bằng tiếng Việt trên điện thoại Android* ” nhằm nghiên cứu tổng quan về xử lý ngôn ngữ tự nhiên và một số phương pháp tổng hợp tiếng nói tiếng Việt từ văn bản đã được ứng dụng và thu được kết quả khả quan, đồng thời xây dựng ứng dụng đọc báo bằng tiếng Việt trên điện thoại Android.

Nội dung của luận văn được trình bày trong ba phần chính như sau:

1. Phần mở đầu

2. Phần nội dung: bao gồm ba chương:

Chương 1: Tổng quan về xử lý ngôn ngữ tự nhiên

Chương 2: Một số phương pháp tổng hợp tiếng nói tiếng Việt

Chương 3: Xây dựng ứng dụng

3. Phần kết luận

Chương 1. TỔNG QUAN VỀ XỬ LÝ NGÔN NGỮ TỰ NHIÊN

1.1. Giới thiệu về xử lý ngôn ngữ tự nhiên

1.1.1. Ngôn ngữ

Ngôn ngữ được coi làm một hệ thống giao thiệp hay suy luận. Hệ thống này dùng một cách biểu diễn phép ẩn dụ và một loại ngữ pháp theo logic, mỗi thứ đều bao hàm một tiêu chuẩn hay sự thật thuộc lịch sử và siêu việt. Hầu hết các ngôn ngữ sử dụng điệu bộ, âm thanh, ký hiệu hay chữ viết để truyền tải khái niệm, ý nghĩa và ý nghĩ nhưng nhiều khi những khía cạnh này khá là giống nhau nên rất khó phân biệt [3].

1.1.2. Xử lý ngôn ngữ tự nhiên

Xử lý ngôn ngữ tự nhiên (Natural language processing- NPL) là một nhánh của trí tuệ nhân tạo tập trung vào các ứng dụng trên ngôn ngữ của con người. Trong trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó liên quan đến việc phải hiểu ý nghĩa của ngôn ngữ - công cụ hoàn hảo nhất của tư duy và giao tiếp.

Xử lý ngôn ngữ tự nhiên là một lĩnh vực nghiên cứu nhằm giúp cho các hệ thống máy tính hiểu và xử lý được ngôn ngữ con người. Tổng hợp tiếng nói là một trong những ứng dụng chính của xử lý ngôn ngữ tự nhiên. Mặc dù tổng hợp tiếng nói đã được nghiên cứu và phát triển trong nhiều năm qua, song vẫn tồn tại nhiều vấn đề cần nghiên cứu.

1.2. Chuẩn hóa văn bản

1.2.1. Tổng quan về chuẩn hóa văn bản

Trong lĩnh vực ngôn ngữ và công nghệ liên quan tới tiếng nói nói chung theo cách này hay cách khác đều phải giải quyết bài toán về xử lý văn bản trong thực tế. Một số lĩnh vực phụ thuộc trực tiếp vào việc giải quyết bài toán này, như máy dịch ngôn ngữ, hệ thống phát hiện chủ đề văn bản, hệ thống tổng hợp tiếng nói từ văn bản. Một số lĩnh vực lại phụ thuộc gián tiếp như nhận dạng tiếng nói sử dụng mô hình ngôn ngữ, trong khi mô hình ngôn ngữ sử dụng các văn bản làm tập huấn luyện. Trong trường hợp nào đi nữa thì đều phải đối mặt với các vấn đề của văn bản thực tế, đó là tính hỗn độn của văn bản.

Chuẩn hóa văn bản thực chất là đi tìm từ điển giải tương ứng để có thể áp dụng được luật phiên âm cho từ chưa chuẩn hóa, từ tương ứng đó chỉ ra cách đọc cho từ chưa chuẩn hóa.

1.2.2. Các nghiên cứu liên quan trên thế giới

Trên thế giới đã có nhiều kết quả nghiên cứu về chuẩn hóa văn bản ở các ngôn ngữ khác nhau, như tiếng Anh [11], Hindi [8], Bangla [4], Trung [10] [14] ... và đã đạt được nhiều thành tựu, giải quyết một số bài toán đặc thù cho loại ngôn ngữ mà nghiên cứu đó tập trung.

1.2.3. Các nghiên cứu liên quan cho tiếng Việt

Ở Việt Nam hiện nay, đề tài về xây dựng bộ tổng hợp tiếng nói cho tiếng Việt cũng đã được quan tâm nghiên cứu, nhiều nghiên cứu đã gặt hái những thành quả đầu tiên trong lĩnh vực này như bộ tổng hợp tiếng nói SAOMAI, HOASUNG, Tiếng Nói PHƯƠNG NAM... Nhưng những nghiên cứu này chưa chú trọng nhiều vào chuẩn hóa văn bản mà chủ yếu tập trung vào việc xử lý tín hiệu. Một số khác xoay quanh bài toán chỉnh sửa lại chính tả. Vì thế dù chất lượng tiếng nói tổng hợp ra khá tốt, nhưng những bộ tổng hợp tiếng nói này chỉ có khả năng làm việc tốt với những văn bản đầu vào có định dạng đơn giản và tương đối chuẩn.

1.2.4. Chuẩn hóa văn bản tiếng việt

Văn bản tiếng Việt thường hàm chứa những dạng chữ số (số đếm, số điện thoại, thời gian...), những tổ hợp chữ có số (kí hiệu, mã ...), những loại dấu, từ viết tắt (TS, Ths...), kí hiệu, từ mượn (FAO, WHO, NATO...) [11] ... chính là những từ chưa chuẩn hóa hay Non-Standard Word (NSW). Việc chuẩn hóa văn bản là để diễn giải những NSW này để bộ tổng hợp tiếng nói có thể hiểu được.

Văn bản tiếng Việt ngoài những vấn đề chung của bài toán chuẩn hóa văn bản còn có những yếu tố đặc thù riêng của nó. Đó là sự nhập nhằng khá phổ biến xảy ra trong các văn bản và cách viết, cách đọc của từng người nhiều khi rất đa dạng, thậm chí không theo quy chuẩn nào [17] [18].

1.3. Phân tích cú pháp

1.3.1. Tổng quan về phân tích cú pháp

Trong tổng hợp tiếng nói, phân tích cú pháp đóng một vai trò rất quan trọng trong công đoạn xử lí văn bản của hệ thống. Phân tích cú pháp chuẩn xác sẽ đưa ra cho hệ thống một cái nhìn toàn cảnh về cấu trúc của văn bản, các cụm từ trong văn bản từ phức tạp đến đơn giản nhất, đồng thời các vị trí âm tiết trong cụm từ cũng được đưa ra luôn.

Phân tích cú pháp là nhằm phân tích một câu thành những thành phần văn phạm có liên quan với nhau và được thể hiện thành cây cú pháp. Khi nhập câu, ta phải phân thành các

thành phần như: chủ ngữ, vị ngữ; gán vai trò chủ từ/đối từ của động từ chính, bổ nghĩa,.. Để phân tích cú pháp, chúng ta cần có bộ luật văn phạm và giải thuật phân tích cú pháp.

1.3.2. Các nghiên cứu về phân tích cú pháp

Trên thế giới, bài toán phân tích cú pháp đã được nghiên cứu và triển khai từ rất lâu. Đặc biệt với tiếng Anh, đã có rất nhiều thành công và đã tiến rất xa. Các mô hình PCFG (Probabilistic context-free grammar), HPCFG (Head-lexicalised probabilistic context-free grammar)... đã cho kết quả phân tích cú pháp rất khả quan.

Tại Việt Nam, những kết quả nghiên cứu về phân tích cú pháp tiếng Việt rất ít và nếu có thì cũng không được phổ biến rộng rãi. Kết quả nghiên cứu rất khả quan nhưng đã cách đây khá lâu (1990 và 1998). Tập luật xây dựng được đưa ra cũng chưa phải là đầy đủ và cũng không thể tạo điều kiện tốt cho bước phân tích ngữ nghĩa tiếp sau [1].

1.4. Phân tích ngữ cảnh

Mục đích của việc phân tích ngữ cảnh là kiểm tra ý nghĩa của câu có mâu thuẫn với ý nghĩa của đoạn hay không. Dựa trên mối liên hệ logic về nghĩa giữa các cụm từ trong câu và mối liên hệ giữa các câu trong đoạn, hệ thống sẽ xác định được (một phần) ý nghĩa của câu trong ngữ cảnh của đoạn.

1.4.1. Nhập nhằng nghĩa ở mức từ vựng

Xét ví dụ “Tôi với quả cam ở trên cây”, ta có từ “với” là “liên từ” hoặc “động từ”. Để chọn được nghĩa thích cho từ “với” trong trường hợp này chúng ta phải vận dụng các ý niệm của ngôn ngữ học tri nhận để biết rằng “với” là động từ chỉ hành động tác động đến một danh từ chỉ sự vật”, và “với” là liên từ liên kết giữa hai đối tượng có cùng kiểu”. Kết hợp những ý niệm ấy, ta có “tôi” là đại từ và “quả cam” là danh từ chỉ sự vật không thuộc cùng dạng đối tượng, do đó máy tính sẽ chọn từ “với” có nghĩa là “Động từ” cho trường hợp này.

1.4.2. Mức độ nhập nhằng cấu trúc

Ví dụ xét câu “*Một con sói và một bầy cừu non*”, ta có 2 phân tích: “[Một con sói] và [một bầy cừu non]” và “[Một con sói và một bầy cừu] non”, máy tính sẽ chọn cách phân tích thứ hai (do tính cân bằng vốn có trong cấu trúc song song của liên từ “và”). Tuy nhiên, nếu xét “*Một đứa trẻ và một người đàn ông già*”, ta cũng sẽ có 2 phân tích: “[đứa trẻ] và [người đàn ông già]” và “[đứa trẻ và người đàn ông] già” và máy tính sẽ chọn cách phân tích thứ

nhất, vì máy thấy cấu trúc thứ hai là vô lý (do có sự đối lập về ngữ nghĩa giữa thuộc tính “trẻ” trong “đứa trẻ” và thuộc tính “già” trong “người đàn ông”).

1.4.3. Mức độ nhập nhằng liên câu

Ví dụ xét câu “*Con cá Sấu săn mồi vì nó đói*”, máy tính hiện nay, trong một số trường hợp, có thể xác định được đại từ “nó” thay thế cho từ nào: “cá Sấu” hay “mồi”. Để giải quyết được nhập nhằng này, máy tính phải xem lại mệnh đề trước và vận dụng tri thức về thế giới thực để biết rằng “chỉ có cá Sấu mới có khả năng đói” nên sẽ chọn “nó thay thế cho cá Sấu”.

1.5. Giới thiệu về hệ thống tổng hợp tiếng nói

1.5.1. Tổng quan

Tổng hợp tiếng nói là việc tạo ra tiếng nói của con người một cách nhân tạo, một hệ thống thực hiện mục đích này được gọi là một hệ thống tổng hợp tiếng nói. Tổng hợp tiếng nói có thể thực hiện bằng phần mềm trên máy tính, các thiết bị di động hay các hệ nhúng.

Chất lượng của một hệ thống tổng hợp tiếng nói được đánh giá dựa trên độ giống, độ tự nhiên với tiếng nói của con người và khả năng để người nghe có thể hiểu được hết ý nghĩa của văn bản.

1.5.2. Ý nghĩa của tổng hợp tiếng nói

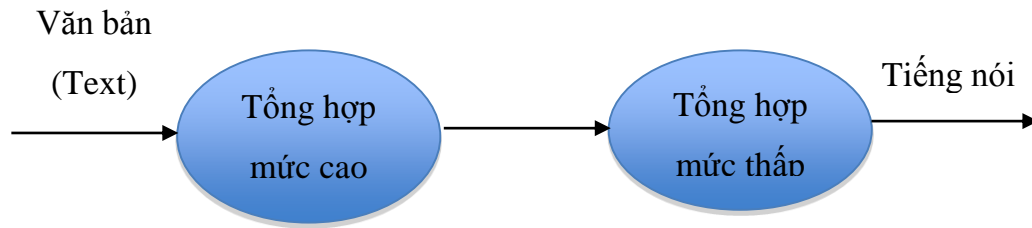
Tổng hợp tiếng nói nói chung và của TTS nói riêng có rất nhiều ý nghĩa thực tiễn. Đặc biệt trên thế giới có nhiều ứng dụng TTS tiếng Anh đã hết sức thành công:

- Giúp đỡ những người bị yếu thị lực, giảm thị lực hoặc tàn tật. Đây là một trong những ý nghĩa to lớn nhất của TTS.
- Ứng dụng trong các thiết bị truyền thông, các nơi công cộng như nhà ga, bệnh viện, sân bay, có cơ quan có hệ thống lấy số xếp hàng.

1.5.3. Mô hình tổng hợp tiếng nói từ văn bản

Thông thường quá trình tổng hợp tiếng TTS nói được chia làm hai mức xử lý:

- Tổng hợp mức cao
- Tổng hợp mức thấp

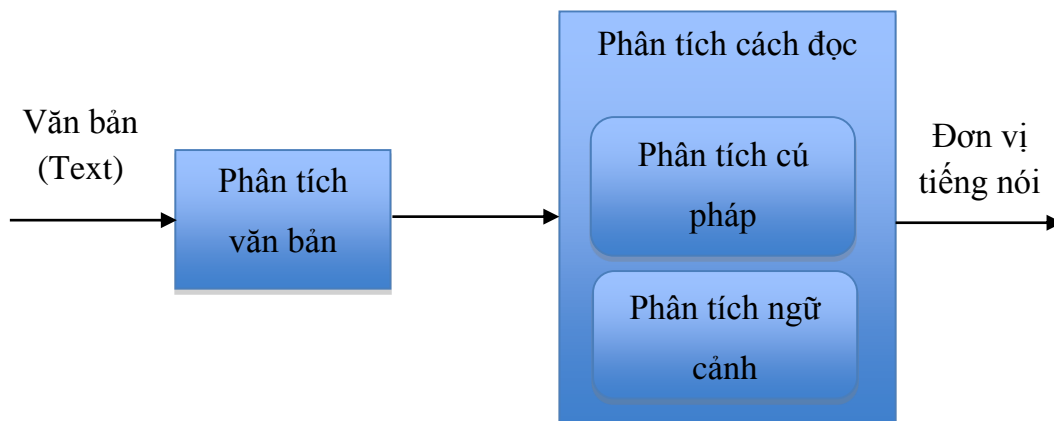


Hình 1.1: Hệ thống tổng hợp tiếng nói

1.5.3.1. Tổng hợp mức cao

Tổng hợp mức cao là ở giai đoạn đầu của quá trình tổng hợp tiếng nói. Ở giai đoạn này sẽ có hai bước chính đó là:

- Chuẩn hóa văn bản
- Phân tích cách đọc



Hình 1.2: Mô hình tổng hợp mức cao

1.5.3.2. Tổng hợp mức thấp

Tổng hợp mức thấp là quá trình kết hợp các đoạn tín hiệu đã được phân tích và xử lý qua quá trình tổng hợp mức cao để tạo ra sóng âm và phát ra tiếng nói. Trên thế giới có nhiều phương pháp được đưa ra để tổng hợp tiếng nói trong giai đoạn này như phương pháp Formant, phương pháp ghép nối diphone,...

Các phương pháp được chia ra năm nhóm chính:

- Phương pháp tổng hợp dựa trên mô phỏng hệ thống phát âm.
- Phương pháp tổng hợp dựa trên hệ luật: phương pháp Formant.

- Phương pháp tổng hợp bằng ghép nối: ghép nối phone, nửa phone, diphone.
- Phương pháp tổng hợp dựa trên các mô hình: mô hình Markov ẩn (HMM).
- Phương pháp tổng hợp dựa trên lai ghép.

Chương 2. MỘT SỐ PHƯƠNG PHÁP TỔNG HỢP TIẾNG NÓI TIẾNG VIỆT

2.1. Tổng hợp mô phỏng hệ thống phát âm

Tổng hợp mô phỏng hệ thống phát âm là phương pháp mà con người cố gắng mô phỏng quá trình tạo ra tiếng nói sao cho càng giống với cơ chế phát âm của con người càng tốt.

2.1.1. *Hệ thống tiếng nói con người*

2.1.1.1. Bộ máy phát âm

Bộ máy phát âm bao gồm các thành phần riêng rẽ như phổi, khí quản, thanh quản, và các đường dẫn miệng, mũi.

2.1.1.2. Cơ chế phát âm

Tiếng nói được tạo ra do tín hiệu nguồn từ thanh môn phát ra, đẩy không khí có trong phổi lên tạo thành dòng khí, va chạm vào hai dây thanh trong tuyến âm. Hai dây thanh dao động sẽ tạo ra cộng hưởng, dao động âm sẽ được lan truyền theo tuyến âm (tính từ tuyến âm đến khoang miệng) và sau khi đi qua khoang mũi và môi, sẽ tạo ra tiếng nói.

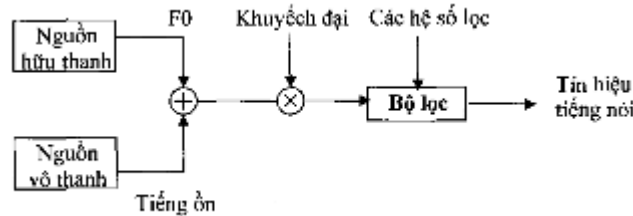
2.1.1.3. Hệ thống tổng hợp mô phỏng phát âm

Hệ thống tổng hợp mô phỏng phát âm đầu tiên là ASY, thường được dùng trong các phòng thí nghiệm trong nghiên cứu, được phát triển ở phòng thí nghiệm Haskins vào giữa những năm 1970 bởi Philip Rubin, Tom Baer, và Paul Mermelstein. ASY dựa trên mô hình cơ quan phát âm đã được tạo ra bởi phòng thí nghiệm Bell vào những năm 1960 và 1970 bởi Paul Mermelstein, Cecil Coker, và các đồng nghiệp khác.

Do những hạn chế trong vấn đề mô phỏng các tham số tiếng nói và năng lực tính toán, mà tổng hợp mô phỏng hệ thống phát âm đã không đạt được nhiều thành công mong đợi như phương pháp tổng hợp tiếng nói khác. Tuy nhiên, nó có rất nhiều ứng dụng hữu ích trong nghiên cứu cơ bản về quá trình tạo ra tiếng nói, và hiện nay phương pháp này đang được đầu tư nghiên cứu và phát triển trở lại [15].

2.2. Tổng hợp tần số formant

Tổng hợp tần số formant hay còn gọi là tổng hợp formant là một trong những phương pháp dựa trên lý thuyết âm học của quá trình tạo ra tiếng nói. Mô hình bộ tổng hợp là một hệ thống nguồn gồm nguồn âm và bộ lọc (hình 2.1). Các tần số formant và các tham số đặc trưng khác là tham số điều khiển mô hình này.

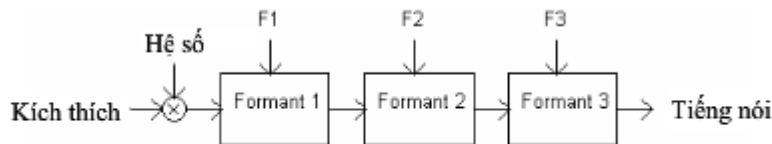


Hình 2.1: Mô hình tổng hợp tần số formant

2.2.1. Các mô hình tổng hợp formant

2.2.1.1. Bộ tổng hợp formant nối tiếp

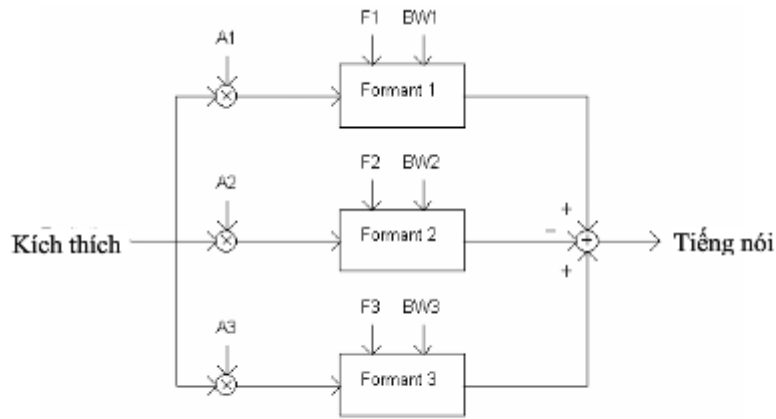
Bộ tổng hợp formant nối tiếp là một bộ tổng hợp formant có các tầng nối tiếp, đầu ra của bộ cộng hưởng này là đầu vào của bộ cộng hưởng kia.



Hình 2.2: Mô hình bộ tổng hợp formant nối tiếp

2.2.1.2. Bộ tổng hợp formant song song

Bộ tổng hợp formant song song bao gồm các bộ cộng hưởng mắc song song. Đầu ra là kết hợp của tín hiệu nguồn và tất cả các formant. Cấu trúc song song cần nhiều thông tin để điều khiển hơn.



Hình 2.3: Mô hình bộ tổng hợp formant song

2.2.2. Tổng hợp tiếng nói trên cơ sở tổng hợp formant

Một trong những hạn chế thường được đề cập đến khi bàn về mô hình tổng hợp formant dựa trên mô hình nguồn âm, bộ lọc là tiếng nói tạo ra nghe như “robot”. Lý do là vì mô hình này mô tả tốt cho âm hữu thanh và các tần số formant nhưng không có đặc trưng vật lý của các tuyến âm. Ưu điểm của mô hình tổng hợp formant là dữ liệu và chương trình rất nhỏ, đặc biệt có thể điều khiển mềm dẻo các thông số đặc trưng của tiếng nói điều này rất quan trọng trong việc xây dựng các hệ thống tổng hợp tiếng nói có chất lượng cao. Mô hình tổng hợp formant mà tiêu biểu nhất là mô hình tổng hợp của Klatt, đã có các sản phẩm thương mại nổi tiếng như DECtalk (tiền thân là MITALK) thành công với mô hình này.

Hiện nay, với những công cụ thích hợp chúng ta hoàn toàn có thể xác định tần số formant cho các âm vị của tiếng Việt [15][16]. Tuy nhiên, phương pháp này vẫn còn tồn tại một số nhược điểm như là khó xây dựng, cần nghiên cứu sâu sắc về ngữ âm của ngôn ngữ, phức tạp trong việc xác định các tham số điều khiển bộ tổng hợp, hạn chế về tính tự nhiên, độ giống tiếng người của tiếng nói tạo ra, chất lượng tiếng nói không tự nhiên (nói nghe như tiếng robot, khác hoàn toàn giọng nói con người) và phụ thuộc nhiều vào chất lượng của quá trình phân tích tiếng nói của từng ngôn ngữ.

2.3. Tổng hợp dựa trên ghép nối

Tổng hợp ghép nối (hay còn gọi là lựa chọn đơn vị âm) là một trong số các phương pháp tổng hợp mới phát triển sau này, kết hợp hay còn gọi là ghép nối các mẫu tiếng nói tự nhiên thu âm sẵn lại với nhau để tạo ra câu nói tổng hợp [14]. Đơn vị âm (unit) phổ biến là âm vị, âm tiết, bán âm tiết, âm đôi, âm ba, từ, cụm từ. Do các đặc tính tự nhiên của tiếng nói

được lưu giữ trong các đơn vị âm, nên tổng hợp ghép nối là phương pháp có khả năng tổng hợp tiếng nói với mức độ dễ hiểu, tự nhiên và có chất lượng cao.

2.3.1. Các vấn đề trong tổng hợp tiếng nói bằng phương pháp ghép nối

- Lựa chọn loại đơn vị âm.
- Xây dựng kho đơn vị âm.
- Tìm kiếm đơn vị âm tối ưu.
- Ghép nối đơn vị âm.

2.3.1.1. Lựa chọn đơn vị âm

- *Âm vị* là loại đơn vị nhỏ nhất trong hệ thống các đơn vị của ngôn ngữ.
- *Âm vị kép* là một đoạn tín hiệu cấu thành từ nửa cuối một đơn vị âm và nửa đầu đơn vị âm tiếp theo.
- *Bán âm tiết* là một phân đoạn tín hiệu của một nửa đầu và nửa cuối của một âm tiết.
- *Âm đầu* là phần phụ âm bắt đầu một âm tiết, phần này là tùy chọn và không mang thông tin về thanh điệu.
- *Vần* là sự kết hợp của ba thành phần: âm đệm, âm chính và âm cuối.
- *Âm tiết* là đơn vị phát âm nhỏ nhất của lời nói, mang những sự kiện ngôn điệu như thanh điệu, trọng âm.
- *Cụm từ* có thể là một hoặc bất kì một đơn vị âm nào

2.3.1.2. Xây dựng kho đơn vị âm

Để xây dựng kho đơn vị âm, các việc cơ bản cần làm là ghi âm các đoạn tiếng nói từ một người thu âm duy nhất và gán nhãn các đoạn tiếng nói với văn bản tương ứng.

Sau khi thu âm dữ liệu văn bản, việc tiếp theo là phân đoạn tín hiệu thành các đoạn tương ứng với đơn vị âm. Quá trình phân đoạn có thể thực hiện tự động hoặc thủ công.

Bước tiếp theo là gán nhãn cho đoạn âm thanh. Các thông số liên quan như trường độ, tần số cơ bản, điểm đánh dấu đường biên của tín hiệu cũng được gán cho đơn vị âm.

2.3.1.1. Tìm kiếm đơn vị âm tối ưu

Văn bản đầu vào được phân tích thành chuỗi các đơn vị âm đích. Các đơn vị âm đích này sẽ được dùng để tìm kiếm trong cơ sở dữ liệu. Mục đích của việc tìm kiếm là chọn ra chuỗi đơn vị tối ưu khớp với ngữ điệu mong muốn nhất.

Hai phương pháp được dùng để lựa chọn các đơn vị âm tối ưu là:

- Chọn lựa dựa trên mô hình cây quyết định
- Chọn lựa dựa trên việc tối ưu hóa hàm chi phí

2.3.2. *Các phương pháp tổng hợp bằng ghép nối*

2.3.2.1. Phương pháp tổng hợp chọn đơn vị

Tổng hợp chọn đơn vị sử dụng một cơ sở dữ liệu lớn các giọng nói ghi âm. Trong lúc ghi âm, mỗi câu phát biểu được tách ra thành các đơn vị âm khác nhau như: các tiếng đơn lẻ, phone, từ, nhóm từ hoặc câu văn.

Thông thường, việc tách ra như vậy cần một máy nhận dạng tiếng nói được đặt ở chế độ so khớp với văn bản viết tương ứng với đoạn ghi âm và dùng đến hiển thị sóng âm và phổ âm thanh. Một bảng tra các đơn vị được lập ra dựa trên các phần đã tách và các thông số âm học như tần số cơ bản, thời lượng, vị trí của âm tiết và các tiếng gần đó. Khi chạy, các câu phát biểu được tạo ra bằng cách xác định chuỗi đơn vị phù hợp nhất từ cơ sở dữ liệu. Quá trình này được gọi là chọn đơn vị, và thường cần dùng đến cây quyết định để thực hiện.

Kỹ thuật chọn đơn vị tạo ra tiếng nói có chất lượng và độ tự nhiên cao do không áp dụng các kỹ thuật xử lý tín hiệu số lên các đoạn giọng nói đã ghi âm, tuy rằng một số hệ thống có thể áp dụng xử lý tín hiệu tại các đoạn nối giữa các tiếng để làm liền mạch kết quả sau khi ghép nối. Kỹ thuật thường được sử dụng để xử lý tín hiệu tại các điểm nối là PSOLA (Pitch Synchronous Overlap and Add).

2.3.2.2. Phương pháp PSOLA

Phương pháp PSOLA bao gồm 3 bước cơ bản:

- Phân tích tín hiệu thành các sóng cơ bản.
- Tính toán các điểm đánh dấu cao độ: bước này sẽ thực hiện biến đổi trường độ và cao độ của tín hiệu. Việc biến đổi cao độ được thực hiện bằng cách thay đổi khoảng cách giữa các sóng cơ bản thu được ở bước phân tích. Việc biến đổi trường độ tín hiệu được thực hiện bằng việc lặp lại hoặc bỏ bớt các sóng cơ bản. Lặp lại thì sẽ làm tăng trường độ, bỏ bớt làm giảm trường độ.
- Tổng hợp lại các đoạn tín hiệu đã được biến đổi

2.3.2.3. Các phiên bản của PSOLA

- TD-PSOLA (Time Domain - PSOLA)
- FD-PSOLA (Frequency Domain - PSOLA)
- LP-PSOLA (Linear Prediction – PSOLA)

2.3.2.4. Vấn đề không liên tục trong ghép nối

Khi sử dụng kỹ thuật PSOLA cho việc ghép nối các đơn vị âm, sẽ vẫn tồn tại ba khả năng về sự không liên tục có thể xảy ra: không liên tục về pha, tần số cơ bản và phổ [5] .

Sự không liên tục về pha: xảy ra do có sự khác nhau về vị trí của các điểm đánh dấu cao độ giữa các đoạn tín hiệu trái và phải.

Sự không liên tục về tần số cơ bản: xảy ra do các đoạn tín hiệu ghép nối có các tần số cơ bản khác nhau.

Sự không liên tục về phổ: xảy ra do hiện tượng đồng cấu âm, gây ra những ảnh hưởng khác nhau lên các đoạn tín hiệu tiếng nói phía trái và phía phải mà xuất phát từ ngữ cảnh khác nhau.

2.3.3. Tổng hợp chuyên biệt

Tổng hợp chuyên biệt ghép nối các từ, và đoạn văn đã được ghi âm để tạo ra lời phát biểu. Nó được dùng trong các ứng dụng có các văn bản chuyên biệt cho một chuyên ngành, sử dụng lượng từ vựng hạn chế, như các thông báo chuyên bay hay dự báo thời tiết.

2.4. Tổng hợp dùng tham số thống kê

2.4.1. Tổng quan về tổng hợp dùng tham số thống kê

Tổng hợp tiếng nói sử dụng các HMM (Hidden Markov Model) [9], [12], [13], [16] cũng là một trong các phương pháp được nghiên cứu rộng rãi hiện nay. Ở đây, HMM là mô hình thống kê, sử dụng để mô hình hoá các tham số tiếng nói của một đơn vị ngữ âm, trong một ngữ cảnh cụ thể, được trích rút đồng thời từ cơ sở dữ liệu tiếng nói. Nhờ tập các HMM này, hệ thống sau đó có thể phát sinh ra các tham số tiếng nói, tùy thuộc vào nội dung văn bản đầu vào, để tạo ra tiếng nói dưới dạng sóng nhờ các tham số được phát xạ này.

2.4.2. Mô hình Markov ẩn

Mô hình Markov ẩn được mở rộng khái niệm từ mô hình Markov bằng cách mỗi trạng thái được gắn với một hàm phát xạ quan sát (observation distribution). Ngoài quá trình ngẫu

nhiên chuyển giữa các trạng thái, tại mỗi trạng thái còn có một quá trình ngẫu nhiên sinh ra một quan sát. Như vậy trong Mô hình Markov ẩn có một quá trình ngẫu nhiên kép, trong đó có một quá trình ngẫu nhiên không quan sát được. Tập các quan sát O được sinh ra bởi dãy các trạng thái S_1, S_2, \dots, S_n của mô hình, mà dãy các trạng thái này là không thấy được, đó chính là lý do mô hình được gọi là mô hình Markov ẩn [7].

Nhìn chung một mô hình HMM có thể coi như bộ sinh trạng thái hữu hạn, áp dụng trong nhận dạng tiếng nói thì mỗi dãy trạng thái của mô hình này có thể biểu diễn một âm vị hay một vị trí tương đối tĩnh của cơ quan cấu âm, còn chuỗi quan sát là chuỗi các vector đặc trưng được trích chọn.

2.5. Tổng hợp bằng phương pháp lai ghép

Hệ thống tổng hợp ghép nối dựa trên sự chọn lựa và ghép nối các đơn vị âm thanh thu âm trước. Đây là phương pháp tổng hợp phổ biến nhất do chất lượng tiếng nói tổng hợp rất cao. Tuy nhiên, nhược điểm của nó là chất lượng sẽ bị giảm sút nếu dữ liệu không đủ lớn, thậm chí không thể ghép nối được nếu nội dung cần tổng hợp không có trong cơ sở dữ liệu.

Hệ thống TTS tham số thống kê dựa trên các tham số sinh ra từ tập HMM đã huấn luyện. Hệ thống này có khả năng tạo ra tiếng nói khá mượt mà và khắc phục được các hạn chế của phương pháp ghép nối.

Kết hợp hai phương pháp trên ta được một hệ thống tổng hợp tiếng nói mới là hệ thống tổng hợp tiếng nói bằng phương pháp lai ghép.

Hệ thống tổng hợp tiếng nói bằng phương pháp lai ghép được chia thành hai loại chính:

- Hệ thống tổng hợp lai ghép hướng ghép nối (Concatenation-Oriented)
- Hệ thống tổng hợp lai ghép hướng HMM (HMM-Oriented)

2.5.1. Hệ thống tổng hợp lai ghép hướng ghép nối

Hệ thống tổng hợp lai ghép hướng kết nối là hệ thống tổng hợp tiếng nói sử dụng các HMM để hỗ trợ quá trình ghép. Về cơ bản, hệ thống này thực hiện ghép nối các đơn vị tiếng nói tự nhiên từ các đơn vị đích đã chọn trước thông qua phân cụm dựa trên cây quyết định.

Các vấn đề cần giải quyết trong tổng hợp lai ghép hướng kết nối:

- Dự đoán mục tiêu

- Làm mịn đơn vị
- Hoà trộn đơn vị

2.5.2. Hệ thống tổng hợp lai ghép hướng HMM

Hệ thống tổng hợp lai ghép hướng HMM sử dụng thuật toán tăng cường và hàm trọng số để hoà trộn các tham số HMM với các đơn vị âm thanh tự nhiên. Quá trình tổng hợp không trộn lẫn các đơn vị âm thanh, vì điều này sẽ làm suy giảm chất lượng do tính chất phổ của chúng khác nhau. Ý tưởng của hệ thống là hoà trộn các đoạn (segment) thay vì ghép nối chúng lại. Đoạn ở đây bao gồm các đơn vị tiếng nói tự nhiên (unit) và các chuỗi HMM.

2.5.2.1. Mô hình hoạt động của hệ thống lai ghép hướng HMM

Giống như các hệ thống TTS dựa trên HMM truyền thống, hệ thống lai ghép hướng HMM cũng bao gồm hai giai đoạn: huấn luyện và tổng hợp. Cơ sở dữ liệu tiếng nói chứa các file âm thanh tiếng nói (mỗi file là một câu thu âm) và tập các nhãn tương ứng (chứa thông tin về các phân tử tiếng nói trong các file âm thanh).

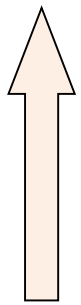
Về mặt chức năng, hệ thống gồm hai mô đun chính:

- *Thành phần dựa trên HMM*: có nhiệm vụ tạo ra các chuỗi tham số bằng cách sử dụng thuật toán sinh tham số.
- *Mô đun ghép nối*: có nhiệm vụ chọn lựa các đơn vị âm thanh tự nhiên từ cơ sở dữ liệu giọng nói đích.

2.6. Đánh giá và lựa chọn phương pháp xây dựng ứng dụng

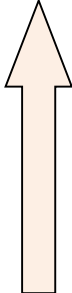
❖ *Về chất lượng của tiếng nói tổng hợp*:

Bảng 2.1: Đánh giá về chất lượng tiếng nói

Phương pháp	Chất lượng
1, Tổng hợp mô phỏng hệ thống phát âm	<div style="display: flex; align-items: center; justify-content: center;"> <div style="margin-right: 10px;">Cao</div>  <div style="margin-left: 10px;">Thấp</div> </div>
2, Tổng hợp bằng phương pháp lai ghép	
3, Tổng hợp dựa trên ghép nối	
4, Tổng hợp dùng tham số thống kê	
5, Tổng hợp tần số formant	

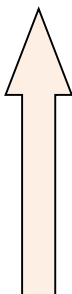
❖ Về hiệu quả tính toán

Bảng 2.2: Đánh giá về hiệu quả tính toán

Phương pháp	Chi phí tính toán
1, Tổng hợp mô phỏng hệ thống phát âm	Cao  Thấp
2, Tổng hợp bằng phương pháp lai ghép	
3, Tổng hợp dùng tham số thống kê	
4, Tổng hợp dựa trên ghép nối	
5, Tổng hợp tần số formant	

❖ Về kích thước dữ liệu

Bảng 2.3: Đánh giá về kích thước dữ liệu

Phương pháp	Chi phí tính toán
1, Tổng hợp dựa trên ghép nối	Cao  Thấp
2, Tổng hợp bằng phương pháp lai ghép	
3, Tổng hợp dùng tham số thống kê	
4, Tổng hợp tần số formant	
5, Tổng hợp mô phỏng hệ thống phát âm	

Với mục đích nghiên cứu tổng hợp tiếng nói tiếng Việt và dựa trên những ưu, nhược điểm của các phương pháp tổng hợp tiếng nói. Luận văn sẽ sử dụng phương pháp tổng hợp bằng ghép nối cho tiếng Việt để xây dựng ứng dụng cho tiếng nói được tổng hợp ra từ phương pháp này.

Chương 3. XÂY DỰNG ỨNG DỤNG

3.1. Giới thiệu về Android SDK

3.1.1. *Android*

Android là một hệ điều hành có mã nguồn mở dựa trên nền tảng Linux được thiết kế dành cho các thiết bị di động có màn hình cảm ứng như điện thoại thông minh và máy tính bảng. Ban đầu, Android được phát triển bởi Tổng công ty Android, với sự hỗ trợ tài chính từ Google, sau này được chính Google mua lại vào năm 2005 và hệ điều hành Android đã ra mắt vào năm 2007.

3.1.2. *Android SDK*

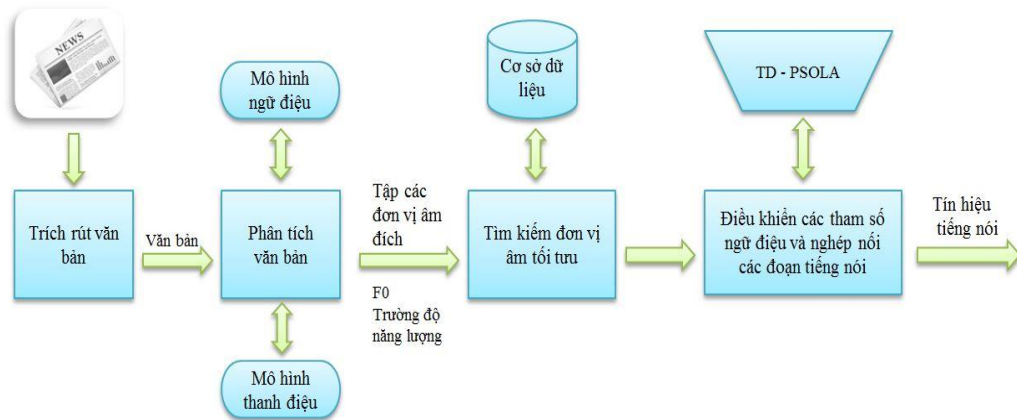
SDK là một thuật ngữ được Microsoft, Sun Microsystems và một số công ty khác sử dụng. Đây là viết tắt của cụm từ Software Development Kit – một bộ công cụ phát triển phần mềm.

Android SDK là một bộ công cụ phát triển ứng dụng cho các thiết bị chạy hệ điều hành Android. Bộ SDK này cung cấp các thư viện API và các công cụ phát triển cần thiết để xây dựng, kiểm tra và các ứng dụng gỡ lỗi cho Android. Trong đó, Text to Speech là một API sẽ được sử dụng trong việc xây dựng ứng dụng.

3.2. Mô tả ứng dụng

3.2.1. *Tổng quan về ứng dụng*

Chương trình đọc báo bằng tiếng Việt trên hệ điều hành Android là chương trình tự động trích rút các thông tin trên các trang báo mạng và dựa trên bộ tổng hợp tiếng nói có sẵn để chuyển hóa thông tin ấy thành lời nói đến người dùng.



Hình 3.1: Mô hình tổng quan về ứng dụng

3.2.2. Tổng quan về giao diện và hoạt động của ứng dụng

3.2.2.1. Sơ đồ Usecase-Actor tổng quan

3.2.2.2. Xây dựng kịch bản

3.2.2.3. Sơ đồ hoạt động

3.2.2.4. Giao diện ứng dụng

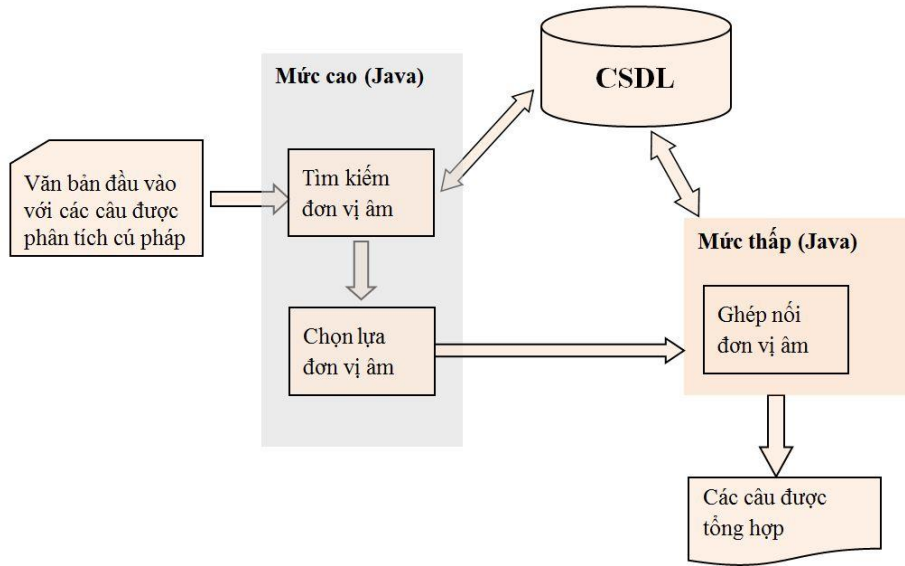
3.3. Tổng hợp tiếng nói từ văn bản trên hệ điều hành Android

3.3.1. Tính năng TextToSpeech trên hệ điều hành Android

Tính năng chuyển văn bản thành giọng nói (Text-to-speech hay TTS) được Google trang bị sẵn cho hệ điều hành Android từ phiên bản 1.6 Donut. Tính năng rất hữu ích trong nhiều trường hợp, đặc biệt đối với các phần mềm từ điển.

3.3.2. Mô hình tổng hợp tiếng nói trên hệ điều hành Android

Bộ tổng hợp tiếng nói trên hệ điều hành Android được viết trên ngôn ngữ là Java và chạy trên hệ điều hành Android. Mô hình bao gồm các phần:



Hình 3.2: Mô hình tổng hợp tiếng nói trên hệ điều hành Android

3.3.3. Lựa chọn và tìm kiếm đơn vị âm

Quá trình lựa chọn đơn vị cũng được chia thành hai bước là tiền lựa chọn và lựa chọn cuối cùng. Tiền lựa chọn là chọn ra các đơn vị âm dài nhất có thể, bước lựa chọn cuối cùng: lựa chọn ra dãy đơn vị âm tốt nhất.

3.3.3.1. Tiền lựa chọn

Văn bản cần tổng hợp sẽ được chia thành các câu để tìm kiếm. Mỗi câu được phân tách thành các cụm từ và âm tiết và tìm kiếm chúng trong CSDL văn bản. Nếu tìm thấy, vị trí tìm thấy và các thông tin về ngữ cảnh và ngữ âm của đơn vị âm tìm thấy được trả về để dùng cho việc tính toán hàm chi phí. Nếu âm tiết không được tìm thấy, âm tiết sẽ được phân tích thành hai bán âm tiết đầu và cuối. Các bán âm tiết này được tìm kiếm trong CSDL bán âm tiết. Tại mức này hầu như không xảy ra sự kiện không tìm thấy bán âm tiết [5]. Nếu không tìm thấy thì âm tiết đó không được tổng hợp.

3.3.3.2. Lựa chọn cuối cùng

Mục đích của giai đoạn này là chọn ra chuỗi các đơn vị âm sao cho sự không liên tục là nhỏ nhất có thể. Tiêu chí lựa chọn là dựa trên hàm chi phí bao gồm chi phí đích và chi phí ghép nối.

Chi phí ghép nối được tính theo công thức dưới đây:

$$C^c(u_{i-1}, u_i) = \sum_{j=1}^q w_j^c C_j^c(u_{i-1}, u_i)$$

Trong đó: $C_j(u_{i-1}, u_i)$: chi phí ghép nối phụ.

3.4. Vấn đề lưu trữ và xử lý trên thiết bị di động

Hiện nay, trên những điện thoại chạy hệ điều hành Android, bộ nhớ được chia làm 2 loại chính: bộ nhớ trong và bộ nhớ ngoài. Bộ nhớ trong là vùng nhớ khả dụng của thiết bị, có tốc độ truy cập cao nhưng dung lượng bị hạn chế và không thể mở rộng thêm. Bộ nhớ ngoài: là không gian bộ nhớ có thể mở rộng nhưng tốc độ truy cập không cao.

Về tốc độ xử lý của điện thoại thông minh hiện nay đang ngày càng được cải thiện. Với công nghệ đa nhân, nhiều luồng được xử lý cùng một lúc đã cải thiện đáng kể tốc độ xử lý của điện thoại.

3.5. Kết quả và đánh giá ứng dụng

Ứng dụng đã được xây dựng và cài đặt thành công trên điện thoại chạy hệ điều hành Android. Ứng dụng hỗ trợ người dùng đọc báo từ 6 trang báo mạng:

Giao diện ứng dụng đơn giản, dễ sử dụng và không bị dừng trong quá trình sử dụng ứng dụng. Bộ tổng hợp tiếng nói mà ứng dụng sử dụng để đọc các bài báo mạng được phát triển bởi trung tâm nghiên cứu và phát triển bởi Samsung Việt Nam và trung tâm MICA (Đại học bách khoa Hà Nội)

KẾT LUẬN

1. Kết quả đạt được

Trong quá trình thực hiện luận văn, học viên đã nghiên cứu một số kiến thức về xử lý ngôn ngữ tự nhiên cần thiết cho quá trình tổng hợp tiếng nói như: chuẩn hóa văn bản, các nghiên cứu trong và ngoài nước về chuẩn hóa văn bản, phân tích cú pháp, các nghiên cứu về phân tích cú pháp ở trong nước và nước ngoài, phân tích ngữ cảnh, nghiên cứu vấn đề nhập nhằng về từ vựng, cấu trúc và nhập nhằng liên câu. Dựa trên cơ sở đó, học viên tiếp tục nghiên cứu và trình bày các phương pháp tổng hợp tiếng nói đang được sử dụng và phát triển hiện nay như: Phương pháp tổng hợp mô phỏng hệ thống phát âm, phương pháp tổng hợp tần số formant, phương pháp tổng hợp dựa trên ghép nối, phương pháp tổng hợp dùng tham số thống kê và phương pháp tổng hợp bằng phương pháp lai ghép. Sau khi nghiên cứu về các phương pháp tổng hợp trên, học viên tiến hành đánh giá và nhận xét về các phương pháp, chỉ ra cụ thể những ưu điểm và nhược điểm của từng phương pháp. Từ đó, học viên lựa chọn một phương pháp khả thi là phương pháp tổng hợp dựa trên ghép nối để xây dựng ứng dụng đọc báo bằng tiếng Việt trên hệ điều hành Android.

Về mặt ứng dụng, học viên đã xây dựng thành công phần mềm đọc báo bằng tiếng Việt trên điện thoại di động chạy hệ điều hành Android. Trong quá trình xây dựng ứng dụng, học viên có trình bày một số kiến thức về Android liên quan như: Android SDK và tính năng TextToSpeech. Bên cạnh đó, học viên cũng trình bày các bước phân tích thiết kế hệ thống như: xây dựng sơ đồ Usecase – Actor tổng quan, xây dựng kịch bản và sơ đồ hoạt động. Sau khi xây dựng thành công ứng dụng, học viên tiến hành nhận xét và đánh giá về ứng dụng và trình bày cụ thể trong luận văn.

2. Những điểm còn hạn chế

- Chưa tổng hợp được một bộ tổng hợp tiếng nói riêng cho ứng dụng.
- Chưa có cơ hội thử nghiệm trên nhiều người dùng và lấy ý kiến đánh giá cho ứng dụng.
- Tiếng nói phát ra đôi lúc chưa phù hợp với việc đọc bài báo.

3. Hướng phát triển tiếp theo

- Nghiên cứu các phương pháp chuẩn hóa văn bản tiếng Việt để làm giảm độ nhập nhằng ngữ nghĩa trong xử lý văn bản đầu vào.
- Nghiên cứu các phương pháp tóm tắt văn bản để tóm tắt các bài báo mạng, từ đó đọc tóm tắt bài báo .

- Bổ xung phần lọc dữ liệu được tải về từ các trang báo để loại bỏ thông tin dư thừa, không có ích cho người nghe.
- Nghiên cứu, cải tiến webview trong Android để bôi đen các câu chữ đang được đọc trong bài báo.
- Bổ xung phần chuyển văn bản tiếng Việt không dấu thành có dấu để có thể đọc mọi loại văn.

TÀI LIỆU THAM KHẢO

Tài liệu tiếng Việt

- [1] Lê Thanh Hương (2000), “*Phân tích cú pháp tiếng Việt*” – luận văn thạc sỹ, ĐHBK Hà Nội.
- [2] Phạm Thanh Sơn (2014), “*Một số vấn đề về tổng hợp tiếng nói tiếng Việt*”, Khoa CNTT, Đại học thông tin liên lạc Nha Trang.
- [3] Nguyễn Văn Thành (2014), “*Tìm hiểu về xử lý ngôn ngữ tự nhiên và máy dịch, viết chương trình mô phỏng từ điển Việt-Anh*” Đại học bách khoa Hà Nội.

Tài liệu tiếng Anh

- [4] Firoj Alam, S. M. Murtoza Habib, Mumit Khan (2009), “*Text Normalization system for Bangla*”, BRAC University, Bangladesh.
- [5] Tran Do Dat (2007), “*Synthèse de la parole a partir du texte en langue Vietnamiennne*”, Ph.D. Thesis, Thèse en cotutelle internationale MICA, Hanoi.
- [6] Do Van Thao, Tran Do Dat, Nguyen Thi Thu Trang (2013), *non-uniform unit selection in Vietnamese speech Synthesis*, proceeding of the 2nd In 8th ISCA Speech Synthesis Workshop, Barcelona, Spain.
- [7] Minghui Dong, Kim-Teng Lua, Haizhou Li (2006), “*A Unit Selection-based Speech Synthesis Approach for Mandarin Chinese*”, Institute for Infocomm Research.
- [8] Hewlett (2009), “*Hindi Text Normalization*”, Packard Labs Indian.
- [9] Kim, Sang-Jin (2007), “*HMM-Based Korean Speech Synthesizer with Two-Band Mixed Excitation Model for Embedded Applications*”, Doctoral Dissertation, Information and Communications University, Korea.
- [10] Craig Olinsky and Alan W Black (2000), “*Non – Standard Word and Homograph Resolution for Asian Language Text Analysis*”, Language Technologies Institute Carnegie Mellon University.
- [11] Richard Sproat, Alan W Black, Stanley Chen, Shankar Kumar, Mari Ostendorf and Chistopher (1999), “*Normalization of Non-Standard Words*”.
- [12] TokudaK, ZenH, Black, AlanW (2002), “*An HMM-based speech synthesis system applied to English*” Proc. in *IEEE Speech Synthesis Workshop*, Santa Monica, USA.
- [13] Vu Tat Thang, Luong Chi Mai và Satoshi, Nakamura (2009), “*An HMM-based Vietnamese Speech Synthesis System*” Proc. in *Oriental COCOSDA*, Urumqi, China, tr. 116-121.
- [14] Yunqing Xia, Kam-Fai Wong, Wenjie Li (2006), “*A Phonetic-Based Approach to Chinese Chat Text Normalization*”, Association for Computational Linguistics.

- [15] Youcef, T và Mohamed, B (2011), *Speech synthesis techniques. A survey*. 7th International Workshop on Systems, Signal Processing and their Applications, Tipaza *Algeria*, tr.67-70.
- [16] Yamagishi, J (2006), “*An Introduction to HMM-Based Speech Synthesis*, *Technical Report*”, Tokyo Institute of Technology, Japan.

Website

- [17] Hải Thụy (2006).“Lộn xộn từ ABC”. [online]. Đường dẫn: <http://tuoitre.vn/tin/giao-duc/20061111/lon-xon-tu-abc/171894.html>, truy cập ngày 19/4/2016.
- [18] Hải Thụy (2007).“Câu chuyện tiếng Việt, có thể chuẩn hóa tiếng Việt”. [online]. Đường dẫn: <http://tuoitre.vn/tin/giao-duc/20070107/cau-chuyen-tieng-viet-co-the-chuan-hoa-tieng-viet/181459.html>, truy cập ngày 20/4/2016.