Báo Cáo Phân Tích Chuyên Sâu về Tiêu Chí và Thuật Toán Chấm Điểm Bài Thi TOEIC Speaking cho Hệ Thống Đánh Giá Tự Động

Phần 1: Kiến Trúc Đánh Giá của Bài Thi TOEIC Speaking: Phân Tích Nền Tảng

1.1. Giới thiệu về Mục đích và Triết lý của Bài thi TOEIC Speaking

Bài thi TOEIC Speaking được thiết kế nhằm mục đích đo lường năng lực sử dụng tiếng Anh nói trong môi trường làm việc quốc tế.[1] Khác với các bài thi học thuật như TOEFL, TOEIC tập trung vào các tình huống giao tiếp thực tế, dựa trên nhiệm vụ (task-based) mà một người đi làm có thể gặp phải. Triết lý này nhấn mạnh tính ứng dụng, do đó các câu hỏi và kịch bản trong bài thi đều mô phỏng các tình huống đời thực.[2]

Một khía cạnh cốt lõi của bài thi là khả năng cung cấp thông tin đánh giá trên một phổ năng lực rộng, từ người mới bắt đầu đến trình độ nâng cao.[3] Điều này có nghĩa là hệ thống chấm điểm phải có khả năng phân loại và phân biệt các cấp độ biểu hiện ngôn ngữ một cách chi tiết, thay vì chỉ đơn thuần là "đạt" hay "không đạt".[6] Thí sinh không bị kiểm tra kiến thức chuyên ngành kinh doanh, và từ vựng yêu cầu không vượt quá phạm vi sử dụng trong các hoạt động công việc và đời sống hàng ngày.[8]

1.2. Giải cấu trúc Định dạng Bài thi Mới nhất

Theo định dạng cập nhật, bài thi TOEIC Speaking bao gồm 11 câu hỏi, được chia thành 5 phần (Parts), và có tổng thời lượng khoảng 20 phút.[1] Cấu trúc chi tiết của bài thi là một yếu tố nền tảng để xây dựng giao diện và luồng xử lý cho một hệ thống luyện thi.

Một cập nhật quan trọng cần lưu ý là phần thi "Propose a Solution" (Đề xuất giải pháp), vốn là một phần trong các định dạng cũ và có thể vẫn được đề cập trong một số tài liệu [10], đã chính thức bị loại bỏ khỏi bài thi kể từ ngày 7 tháng 8 năm 2021.[9] Việc đảm bảo nền tảng luyện thi tuân thủ định dạng mới nhất này là tối quan trọng để mang lại trải nghiệm sát với thực tế.

Cấu trúc bài thi được thể hiện chi tiết trong bảng dưới đây, tổng hợp từ các nguồn tài liệu chính thức của ETS và các đơn vị khảo thí uy tín.

Bảng 1: Cấu trúc Bài thi TOEIC Speaking (Định dạng Mới nhất)

Phần (Part)	Câu hỏi	Nhiệm vụ (Task)	Thời gian Chuẩn bị	Thời gian Trả lời	Điểm Thô Tối đa
1	1–2	Read a text aloud (Đọc to một đoạn văn)	45 giây/câu	45 giây/câu	3
2	3–4	Describe a picture (Mô tả một bức tranh)	45 giây/câu	30 giây/câu	3
3	5–7	Respond to questions (Trả lời câu hỏi)	3 giây/câu	15 giây (Q5, Q6); 30 giây (Q7)	3
4	8–10	Respond to questions using information provided (Trả lời câu hỏi sử dụng thông tin	45 giây đọc thông tin; 3 giây/câu	15 giây (Q8, Q9); 30 giây (Q10)	3

		cho sẵn)			
5	11	Express an opinion (Bày tỏ quan điểm)	45 giây	60 giây	5

Dữ liệu được tổng hợp từ các nguồn: [2]

Cấu trúc này không phải là một sự sắp xếp ngẫu nhiên mà là một "giàn giáo" kỹ năng được thiết kế có chủ đích. Bài thi bắt đầu với một nhiệm vụ thuần túy về mặt cơ học phát âm (đọc to) và tăng dần độ phức tạp bằng cách thêm vào các lớp kỹ năng: mô tả (Phần 2), phản xạ tự phát (Phần 3), tổng hợp thông tin (Phần 4), và cuối cùng là lập luận (Phần 5). Phần 1 chỉ đánh giá Pronunciation (Phát âm) và Intonation and Stress (Ngữ điệu và Trọng âm).[11] Phần 2 bổ sung Grammar (Ngữ pháp), Vocabulary (Từ vựng), và Cohesion (Sự mạch lạc). Phần 3 tiếp tục thêm vào Relevance (Sự liên quan) và Completeness (Sự đầy đủ). Sự tiến triển này cho thấy bài thi được xây dựng để tạo ra một bức tranh toàn diện về năng lực của người nói, từ những kỹ năng nền tảng nhất đến những kỹ năng phức tạp nhất. Do đó, một hệ thống chấm điểm tự động phải phản ánh được cấu trúc này bằng cách kích hoạt các mô-đun đánh giá khác nhau cho từng phần, thay vì áp dụng một bộ tiêu chí duy nhất cho tất cả các câu trả lời.

1.3. Cơ chế Chấm điểm: Từ Điểm Thô đến Thang điểm Quy đổi

Hệ thống chấm điểm của TOEIC Speaking hoạt động theo cơ chế hai cấp:

1. **Điểm Thô (Raw Scores):** Mỗi câu trả lời được các giám khảo con người (human raters) đã qua đào tạo chấm điểm dựa trên một thang điểm nhất định. Cụ thể, các câu hỏi từ 1 đến 10 (trong định dạng mới là 1-4 và 5-10) được chấm trên thang điểm từ 0 đến 3. Riêng câu hỏi 11 (trước đây có thể là Q10-11 ở một số tài liệu cũ) được chấm trên thang điểm từ 0 đến 5.[9]
2. **Điểm Quy đổi (Scaled Score):** Tổng của các điểm thô này sau đó được chuyển đổi sang một thang điểm cuối cùng từ 0 đến 200.[1]

Một điểm đặc biệt quan trọng trong cấu trúc chấm điểm là sự khác biệt về thang điểm. Trong khi hầu hết các câu hỏi được chấm trên thang điểm 0-3, câu hỏi 11 (Bày tỏ ý kiến) lại được chấm trên thang điểm 0-5.[9] Sự chênh lệch 67% về trọng số điểm thô này không phải là ngẫu nhiên. Nó cho thấy ETS đặt trọng tâm đặc biệt vào câu hỏi này như một bài kiểm tra tổng hợp, đỉnh cao, nhằm đánh giá khả năng tạo ra một bài nói có sự kết nối, kéo dài và mạch lạc (connected, sustained discourse)—một kỹ năng phân biệt rõ rệt các cấp độ năng lực cao

hơn.[19] Do đó, thuật toán chấm điểm tự động phải có trọng số cao hơn đáng kể cho hiệu suất ở nhiệm vụ cuối cùng này.

Điểm quy đổi 0-200 này tương ứng với 8 cấp độ năng lực (Proficiency Levels).[16] Các mô tả chi tiết cho từng cấp độ này, do ETS cung cấp, là kim chỉ nam để hiểu được năng lực ngôn ngữ tương ứng với từng khoảng điểm. Ví dụ:

- **Level 8 (190–200):** Có khả năng tạo ra các bài nói có sự kết nối, kéo dài, phù hợp với môi trường làm việc. Lời nói rất dễ hiểu. Sử dụng tốt ngữ pháp từ cơ bản đến phức tạp, từ vựng chính xác và tinh tế.
- **Level 7 (160–180):** Có thể diễn đạt ý kiến hoặc trả lời các yêu cầu phức tạp một cách hiệu quả. Có thể mắc một số lỗi nhỏ về phát âm, ngữ điệu, ngữ pháp phức tạp hoặc từ vựng không chính xác, nhưng không ảnh hưởng đến thông điệp.
- **Level 6 (130–150):** Có thể tạo ra câu trả lời liên quan, nhưng lý do hoặc giải thích đôi khi không rõ ràng do phát âm, lỗi ngữ pháp hoặc vốn từ vựng hạn chế.
- **Level 5 (110–120):** Thành công hạn chế trong việc bày tỏ quan điểm. Câu trả lời có các vấn đề như ngôn ngữ không chính xác, mơ hồ, lặp lại; ngập ngừng thường xuyên; và ý tưởng hạn chế.
- *Các cấp độ thấp hơn (1-4) cho thấy sự khó khăn ngày càng tăng trong việc hoàn thành các nhiệm vụ cơ bản.* [19]

Cần lưu ý rằng công thức chuyển đổi chính xác từ tổng điểm thô sang điểm quy đổi không được ETS công bố công khai.[24] Điều này đòi hỏi hệ thống tự động phải xây dựng một mô hình chuyển đổi riêng, dựa trên việc hiệu chỉnh với dữ liệu mẫu đã được chấm điểm để mô phỏng gần nhất với thực tế.

Phần 2: Phân Tích Đa Chiều các Tiêu Chí Đánh Giá Cốt Lõi

Mặc dù ETS không công bố rubric chi tiết cho từng phần của bài thi TOEIC Speaking, việc phân tích các tài liệu hướng dẫn, bài thi mẫu và mô tả điểm số cho phép chúng ta xác định các tiêu chí đánh giá cốt lõi. Để làm sâu sắc thêm sự hiểu biết này, chúng ta có thể tham chiếu đến các rubric minh bạch hơn của các bài thi uy tín khác như TOEFL iBT và IELTS, vốn cũng được phát triển dựa trên các tiêu chuẩn tâm lý trắc lượng học nghiêm ngặt.

2.1. Giải cấu trúc các Tiêu chí Chính thức của TOEIC

Các tài liệu chính thức của TOEIC cho thấy một bộ tiêu chí được áp dụng tăng dần qua các phần thi [11]:

1. **Pronunciation (Phát âm):** Khả năng phát âm các âm riêng lẻ một cách rõ ràng, chính xác theo chuẩn quốc tế, bao gồm cả các từ khó mà không ảnh hưởng đến tốc độ nói.[21]
2. **Intonation and Stress (Ngữ điệu và Trọng âm):** Khả năng đặt trọng âm từ và trọng âm câu một cách tự nhiên, sử dụng ngữ điệu lên xuống để truyền tải ý nghĩa và nhấn mạnh thông tin quan trọng.[21]
3. **Grammar (Ngữ pháp):** Khả năng sử dụng các cấu trúc ngữ pháp một cách chính xác và phù hợp để truyền đạt ý tưởng rõ ràng, súc tích.[21]
4. **Vocabulary (Từ vựng):** Khả năng sử dụng vốn từ vựng đa dạng, chính xác và phù hợp với ngữ cảnh để mô tả, giải thích và lập luận.[21]
5. **Cohesion (Sự mạch lạc):** Khả năng liên kết các ý tưởng một cách logic và trôi chảy.
6. **Relevance of content (Sự liên quan của nội dung):** Khả năng trả lời đúng trọng tâm câu hỏi, cung cấp thông tin phù hợp với yêu cầu.
7. **Completeness of content (Sự đầy đủ của nội dung):** Khả năng cung cấp một câu trả lời hoàn chỉnh, không bỏ sót các thông tin quan trọng được yêu cầu.

2.2. Bổ sung Chi tiết từ Rubric của TOEFL iBT và IELTS

Để xây dựng một mô hình chấm điểm tự động tinh vi, việc chỉ dựa vào danh sách trên là không đủ. Việc tham chiếu đến các rubric của TOEFL iBT và IELTS cung cấp một bộ từ vựng và các cấp độ chi tiết hơn để định lượng hóa các tiêu chí này.

Từ Khung Đánh giá của TOEFL iBT [26]:

- **Delivery (Diễn đạt):** Tiêu chí này tương ứng với Pronunciation và Intonation/Stress của TOEIC, đồng thời bổ sung một khái niệm quan trọng là Pacing/Flow (Nhịp độ/Sự trôi chảy). Rubric của TOEFL phân biệt rõ ràng giữa "nhịp điệu/tốc độ bị ngắt quãng" (điểm 2) và "luồng nói nhìn chung có nhịp độ tốt" (điểm 4).
- **Language Use (Sử dụng Ngôn ngữ):** Tương ứng với Grammar và Vocabulary của TOEIC. Rubric này cung cấp sự khác biệt tinh tế giữa "phạm vi và khả năng kiểm soát hạn chế" (điểm 2) và "sử dụng ngữ pháp và từ vựng hiệu quả" với "mức độ tự động hóa cao" (điểm 4).
- **Topic Development (Phát triển Chủ đề):** Tương ứng với Cohesion, Relevance, và Completeness của TOEIC. Nó làm rõ các khái niệm này bằng các mô tả như "mối quan hệ giữa các ý tưởng rõ ràng" và "phát triển tốt và mạch lạc".

Từ Khung Đánh giá của IELTS [28]:

- **Fluency and Coherence (Độ trôi chảy và Mạch lạc):** Cung cấp một hệ thống từ vựng phong phú để mô tả các vấn đề về độ trôi chảy, ví dụ như "lặp lại, tự sửa lỗi và/hoặc nói chậm để tiếp tục" (Band 5) và các vấn đề về mạch lạc như "lạm dụng một số từ nối nhất định" (Band 5). Điều này trực tiếp làm rõ hơn tiêu chí Cohesion của TOEIC.
- **Lexical Resource (Nguồn từ vựng):** Mở rộng tiêu chí Vocabulary của TOEIC bằng cách giới thiệu các khái niệm như "sử dụng từ vựng ít phổ biến và thành ngữ" (Band 7) và khả năng "diễn giải (paraphrase) hiệu quả" (Band 7).
- **Grammatical Range and Accuracy (Phạm vi và Độ chính xác Ngữ pháp):** Phân biệt rõ ràng giữa việc chỉ sử dụng "các dạng câu cơ bản với độ chính xác hợp lý" (Band 5) và việc sử dụng "hỗn hợp các cấu trúc đơn giản và phức tạp" (Band 6).
- **Pronunciation (Phát âm):** Cung cấp một thang đo về mức độ dễ hiểu (intelligibility), từ "phát âm sai thường xuyên và gây khó khăn cho người nghe" (Band 4) đến mức "dễ hiểu một cách dễ dàng" (Band 9).

Từ việc phân tích chéo này, có thể rút ra những kết luận quan trọng cho việc xây dựng hệ thống. Thứ nhất, tiêu chí "Cohesion" trong TOEIC là một thuật ngữ đơn giản hóa cho một tập hợp các kỹ năng phức tạp được mô tả chi tiết trong rubric của IELTS và TOEFL. Nó không chỉ là việc sử dụng từ nối, mà còn bao gồm luồng logic của ý tưởng, cách báo hiệu mối quan-hệ-giữa-các-câu, và cấu trúc tổng thể của câu trả lời. Một mô hình NLP không thể chỉ đếm các từ như "however" hay "therefore"; nó phải có khả năng đánh giá sự tiến triển logic của diễn ngôn.

Thứ hai, có một sự phân biệt quan trọng giữa "Độ chính xác" (Accuracy) và "Phạm vi/Sự linh hoạt" (Range/Flexibility) cho cả Ngữ pháp và Từ vựng. Một thí sinh đạt điểm cao không chỉ là người tránh được lỗi; họ còn thể hiện được khả năng sử dụng ngôn ngữ một cách đa dạng và linh hoạt. Ví dụ, một người nói ở trình độ thấp có thể tạo ra vài câu đơn hoàn hảo về mặt ngữ pháp ("I like dogs. Dogs are cute."). Một người nói ở trình độ cao sẽ sử dụng kết hợp các cấu trúc đơn và phức, ngay cả khi họ mắc một lỗi nhỏ ("Although some people might prefer cats due to their independence, I've always been drawn to the loyalty that dogs offers.").[26] Do đó, hệ thống tự động phải được thiết kế để không chỉ phạt lỗi mà còn thưởng cho sự phức tạp về cấu trúc và sự đa dạng về từ vựng. Điều này có nghĩa là thuật toán chấm điểm cần một thành phần "thưởng" cho việc thể hiện phạm vi, bên cạnh thành phần "phạt" cho các lỗi.

Phần 3: Rubric Chấm Điểm Chi Tiết theo Từng Phần Thi

Dựa trên việc tổng hợp các tiêu chí cốt lõi và các khung đánh giá tham chiếu, phần này sẽ xây dựng một bộ rubric chấm điểm chi tiết, có tính ứng dụng cao cho từng phần của bài thi TOEIC

Speaking.

3.1. Phần 1 (Câu 1-2: Read a Text Aloud) - Rubric về Kiểm soát Âm vị học

Phần này chỉ tập trung vào khả năng đọc và diễn đạt ngữ âm.

- **Tiêu chí chính:** Pronunciation, Intonation and Stress.[11]
- **Điểm 3 (Cao):** Rất dễ hiểu. Phát âm rõ ràng và chính xác. Ngữ điệu và trọng âm tự nhiên, hiệu quả, giúp làm nổi bật ý nghĩa. Tương ứng với mô tả "HIGH" trong các tài liệu của ETS.[20]
- **Điểm 2 (Trung bình):** Nhìn chung dễ hiểu nhưng có một số lỗi. Có thể có một vài từ phát âm sai hoặc trọng âm/ngữ điệu còn lúng túng, nhưng không gây cản trở đáng kể cho việc hiểu. Tương ứng với mô tả "MEDIUM".[20]
- **Điểm 1 (Thấp):** Nhìn chung không dễ hiểu. Các vấn đề về phát âm, trọng âm và ngữ điệu xuất hiện thường xuyên và nhất quán, đòi hỏi người nghe phải nỗ lực rất nhiều. Tương ứng với mô tả "LOW".[20]
- **Điểm 0:** Không trả lời hoặc trả lời bằng ngôn ngữ khác tiếng Anh.[10]

3.2. Phần 2 (Câu 3-4: Describe a Picture) - Rubric về Diễn đạt Mô tả

Phần này đánh giá khả năng sử dụng ngôn ngữ để mô tả một cách có tổ chức.

- **Tiêu chí chính:** Tất cả các tiêu chí của Phần 1, cộng thêm Grammar, Vocabulary, Cohesion.[11]
- **Điểm 3:** Câu trả lời là một bài mô tả liên quan, có tổ chức tốt. Sử dụng từ vựng mô tả chính xác và phù hợp. Ngữ pháp chính xác với sự kết hợp của các cấu trúc câu. Lời nói trôi chảy và dễ hiểu.
- **Điểm 2:** Câu trả lời có liên quan nhưng có thể chưa đầy đủ hoặc thiếu chi tiết. Từ vựng phù hợp nhưng có thể còn cơ bản hoặc lặp lại. Có thể có lỗi ngữ pháp nhưng không làm che mờ ý nghĩa. Việc diễn đạt có thể đòi hỏi người nghe một chút nỗ lực.
- **Điểm 1:** Câu trả lời chỉ liên quan ở mức tối thiểu hoặc rất không đầy đủ. Từ vựng hạn chế hoặc không chính xác. Lỗi ngữ pháp thường xuyên và làm che mờ ý nghĩa. Cách diễn đạt bị ngắt quãng và khó hiểu.
- **Điểm 0:** Không có câu trả lời liên quan.

3.3. Phần 3 (Câu 5-7: Respond to Questions) - Rubric về Phản xạ Tự phát

Phần này nhấn mạnh khả năng hiểu nhanh và trả lời tức thì một cách phù hợp.

- **Tiêu chí chính:** Tất cả các tiêu chí của Phần 2, cộng thêm Relevance of content, Completeness of content.[11]
- **Điểm 3:** Câu trả lời liên quan trực tiếp, đầy đủ và phù hợp. Ý tưởng được diễn đạt rõ ràng và mạch lạc. Kiểm soát tốt ngữ pháp và từ vựng. Diễn đạt trôi chảy và dễ hiểu.
- **Điểm 2:** Câu trả lời có liên quan nhưng có thể chưa đầy đủ hoặc chưa được phát triển. Sự kết nối giữa các ý tưởng không phải lúc nào cũng rõ ràng. Có một số lỗi ngữ pháp hoặc từ vựng. Cách diễn đạt có thể có những khoảng ngập ngừng đáng chú ý.
- **Điểm 1:** Câu trả lời chỉ liên quan ở mức tối thiểu hoặc không trả lời được câu hỏi. Ý tưởng khó hiểu do những hạn chế nghiêm trọng về ngữ pháp và từ vựng, và/hoặc các vấn đề lớn về cách diễn đạt.
- **Điểm 0:** Không có câu trả lời liên quan.

3.4. Phần 4 (Câu 8-10: Respond to Questions Using Information Provided) - Rubric về Tổng hợp Thông tin

Điểm khác biệt chính của phần này là khả năng trích xuất và diễn giải lại thông tin một cách chính xác.

- **Tiêu chí chính:** Tất cả các tiêu chí của Phần 3.[11]
- **Điểm 3:** Truyền đạt chính xác tất cả thông tin được yêu cầu. Diễn giải (paraphrase) hiệu quả thay vì chỉ đọc lại từ màn hình. Câu trả lời được tổ chức tốt, trôi chảy và đúng ngữ pháp.
- **Điểm 2:** Truyền đạt hầu hết thông tin được yêu cầu nhưng có thể có những điểm không chính xác hoặc thiếu sót nhỏ. Có thể phụ thuộc vào việc đọc trực tiếp từ văn bản. Có một số vấn đề về độ trôi chảy, ngữ pháp hoặc từ vựng.
- **Điểm 1:** Không truyền đạt được thông tin chính hoặc chứa những điểm không chính xác đáng kể. Câu trả lời không có tổ chức và khó hiểu do các vấn đề về ngôn ngữ và cách diễn đạt.
- **Điểm 0:** Không có câu trả lời liên quan.

Một kỹ năng ngầm được kiểm tra trong Phần 4 là khả năng diễn giải (paraphrasing) dưới áp lực thời gian. Mặc dù không được nêu rõ, một câu trả lời điểm cao phải biến đổi thông tin văn

bản thành ngôn ngữ nói tự nhiên. Việc chỉ đọc lại văn bản, dù trôi chảy, cũng không thể hiện được trình độ cao. Các câu hỏi thường được diễn đạt dưới dạng hội thoại (ví dụ: "Could you tell me...") [17], và một câu trả lời tự nhiên đòi hỏi phải diễn giải lại. Rubric của IELTS cũng thưởng điểm cao cho khả năng này.[31] Do đó, mô-đun NLP cho Phần 4 phải so sánh nội dung ngữ nghĩa của câu trả lời với văn bản gốc, đồng thời đo lường sự khác biệt về từ vựng và cấu trúc để thưởng cho việc diễn giải thành công và phạt việc sao chép nguyên văn.

3.5. Phần 5 (Câu 11: Express an Opinion) - Rubric về Diễn đạt Lập luận (Thang điểm 0-5)

Đây là nhiệm vụ phức tạp nhất, tích hợp tất cả các kỹ năng trước đó để đánh giá khả năng lập luận.[11] Thang điểm 0-5 [16] cho phép một rubric chi tiết hơn, được xây dựng bằng cách tổng hợp các mô tả cấp độ cao nhất từ TOEIC, TOEFL và IELTS.

Sự khác biệt giữa điểm 4 và 5 ở câu hỏi này nằm ở chất lượng và chiều sâu của các luận điểm hỗ trợ. Cả hai đều có thể có quan điểm rõ ràng và sử dụng ngôn ngữ tốt, nhưng một câu trả lời ở cấp độ cao nhất sẽ cung cấp các luận điểm được phát triển tốt, cụ thể và có sức thuyết phục. Mô tả cấp độ 8 của TOEIC (190-200) đề cập đến việc tạo ra "bài nói có sự kết nối, kéo dài" với từ vựng "chính xác và tinh tế".[19] Rubric điểm 5 cho phần Viết cũng nhấn mạnh "các giải thích, ví dụ, và/hoặc chi tiết rõ ràng và phù hợp".[34] Điều này cho thấy mô-đun NLP cho câu 11 không chỉ cần kiểm tra sự tồn tại của các ý tưởng hỗ trợ mà còn phải cố gắng đánh giá tính cụ thể và sự phát triển của chúng, có thể thông qua việc phân tích độ dài câu, độ phức tạp và việc sử dụng danh từ cụ thể so với các khái quát hóa trừu tượng.

Bảng 2: Rubric Chấm Điểm Toàn Diện cho Phần 5 (Bày tỏ ý kiến)

Điểm	Mô tả Chung	Diễn đạt (Phát âm, Trôi chảy, Ngữ điệu)	Sử dụng Ngôn ngữ (Ngữ pháp, Từ vựng)	Phát triển Chủ đề (Mạch lạc, Liên quan, Đầy đủ, Hỗ trợ)
5	Quan điểm được phát triển tốt, kéo dài và mạch lạc.	Lời nói rất dễ hiểu và trôi chảy. Ngữ điệu và trọng âm được sử dụng hiệu quả để	Thể hiện sự sử dụng linh hoạt và chính xác các cấu trúc ngữ pháp phức tạp và vốn từ	Quan điểm được nêu rõ ràng và được hỗ trợ đầy đủ bằng các lý do, chi tiết và ví dụ

		truyền tải ý nghĩa tinh tế.	vựng rộng. Lỗi rất hiếm và nhỏ.	liên quan, có sức thuyết phục. Các ý tưởng được liên kết một cách logic và chặt chẽ.
4	Quan điểm rõ ràng và được hỗ trợ đầy đủ.	Lời nói trôi chảy với những khó khăn nhỏ không đáng kể. Nhìn chung dễ hiểu.	Sử dụng hiệu quả ngữ pháp và từ vựng, có thể có một số lỗi nhỏ nhưng không làm che mờ ý nghĩa.	Quan điểm được hỗ trợ bằng các lý do và giải thích, mặc dù có thể chưa được phát triển đầy đủ. Mối quan hệ giữa các ý tưởng hầu hết đều rõ ràng.
3	Quan điểm được nêu ra, nhưng sự hỗ trợ còn hạn chế hoặc không rõ ràng.	Việc diễn đạt đòi hỏi nỗ lực từ người nghe do các vấn đề về phát âm, ngập ngừng hoặc nhịp điệu.	Việc kiểm soát ngữ pháp và từ vựng không nhất quán có thể dẫn đến thiếu rõ ràng. Phạm vi ngôn ngữ có thể bị hạn chế.	Quan điểm có thể được hỗ trợ nhưng các lý do còn chung chung hoặc không liên quan. Sự kết nối giữa các ý tưởng có thể bị che mờ.
2	Câu trả lời còn hạn chế, không nêu được quan điểm rõ ràng hoặc không cung cấp được sự hỗ trợ.	Diễn đạt bị ngắt quãng, rời rạc. Khó khăn về phát âm và ngữ điệu gây cản trở đáng kể cho việc hiểu.	Ngữ pháp và từ vựng bị hạn chế nghiêm trọng, ngăn cản việc diễn đạt ý tưởng.	Nội dung rất cơ bản, lặp lại hoặc không liên quan đến nhiệm vụ.

1	Câu trả lời không thể hiểu được, chỉ bao gồm các từ riêng lẻ, hoặc không nêu được quan điểm.	Lời nói phần lớn không thể hiểu được.	Không thể tạo thành các cấu trúc câu cơ bản.	Không có nội dung liên quan.
0	Không trả lời, hoặc trả lời không bằng tiếng Anh.	-	-	-

Phần 4: Kế Hoạch Kỹ Thuật cho Hệ Thống Chấm Điểm bằng AI

Để hiện thực hóa các rubric trên thành một hệ thống tự động, cần một kế hoạch kỹ thuật chi tiết, kết hợp sức mạnh của dịch vụ nhận dạng giọng nói và các mô hình xử lý ngôn ngữ tự nhiên (NLP).

4.1. Mô-đun 1: Đánh giá Âm vị học và Độ trôi chảy với Azure Speech

Dịch vụ Azure Speech Pronunciation Assessment là công cụ chính để đánh giá khía cạnh "Delivery" (Diễn đạt) của bài nói.[35]

Ánh xạ các tính năng của Azure vào Tiêu chí TOEIC:

- **AccuracyScore (Điểm chính xác phát âm):** Đo lường trực tiếp độ chính xác của các âm vị (phoneme), âm tiết (syllable) và từ. Đây là thước đo cốt lõi cho tiêu chí Pronunciation.[35]
- **FluencyScore (Điểm lưu loát):** Đo lường tốc độ nói và các khoảng lặng, ánh xạ trực tiếp đến khía cạnh Fluency của bài thi.[37]
- **ProsodyScore (Điểm ngữ điệu):** Đánh giá ngữ điệu, trọng âm và nhịp điệu, ánh xạ trực tiếp đến tiêu chí Intonation and Stress.[36]

- **CompletenessScore (Điểm đầy đủ):** Đo lường xem tất cả các từ trong một văn bản cho trước có được đọc hay không, rất hữu ích cho Phần 1.
- **Phát hiện ErrorType (Loại lỗi):** Kết quả JSON trả về xác định các lỗi cụ thể ở cấp độ từ như Mispronunciation (Phát âm sai), Omission (Thiếu từ), và Insertion (Thừa từ).[35] Điều này cho phép cung cấp phản hồi cực kỳ chi tiết và xây dựng một thành phần chấm điểm dựa trên việc trừ điểm lỗi.

Chiến lược triển khai:

- **Đối với Phần 1 (Read a text aloud):** Sử dụng kịch bản "Reading" (đọc có kịch bản) của Azure, cung cấp văn bản của bài đọc làm tham chiếu (reference text).[36] Các chỉ số AccuracyScore, ProsodyScore, và CompletenessScore sẽ là đầu vào chính cho thuật toán chấm điểm.
- **Đối với Phần 2-5:** Sử dụng kịch bản "Speaking" (nói không có kịch bản).[36] Mặc dù không có văn bản tham chiếu để so sánh độ chính xác trực tiếp, dịch vụ vẫn cung cấp các chỉ số giá trị về độ trôi chảy, ngữ điệu và độ rõ ràng của từng âm vị. Văn bản được chuyển đổi từ giọng nói (speech-to-text) sẽ được chuyển đến mô-đun NLP để phân tích sâu hơn.

Kết quả từ Azure Speech được trả về dưới định dạng JSON chi tiết, rất lý tưởng cho việc phân tích và xử lý tự động.[35]

4.2. Mô-đun 2: Phân tích Nội dung và Ngôn ngữ với các Mô hình NLP

Mô-đun này xử lý văn bản đã được chuyển đổi từ Azure để đánh giá các tiêu chí ngoài khả năng của dịch vụ giọng nói. Ý tưởng cốt lõi là sử dụng NLP để định lượng hóa các khái niệm trừu tượng trong rubric.[40]

- **Grammatical Range and Accuracy (Phạm vi và Độ chính xác Ngữ pháp):**
  - **Thước đo:** Mật độ lỗi ngữ pháp.
  - **Phương pháp:** Xử lý văn bản bằng một mô hình sửa lỗi ngữ pháp tiên tiến (ví dụ: mô hình T5 đã được tinh chỉnh hoặc một API chuyên dụng). Tính toán theo công thức: $Mật\_độ\_lỗi = (\text{Số lượng lỗi} / \text{Tổng số từ})$.
  - **Thước đo:** Độ phức tạp cú pháp.
  - **Phương pháp:** Sử dụng công cụ gán nhãn từ loại (POS tagging) và phân tích cú pháp phụ thuộc (dependency parsing) để đo lường tỷ lệ câu phức trên câu đơn, độ dài mệnh đề trung bình, và độ sâu của cây cú pháp.[42] Điều này giúp định lượng hóa "phạm vi" ngữ pháp.
- **Lexical Resource (Nguồn từ vựng):**
  - **Thước đo:** Độ đa dạng từ vựng (Lexical Diversity).
  - **Phương pháp:** Tính toán Tỷ lệ Loại-Từ/Tổng-Từ (Type-Token Ratio - TTR) được chuẩn hóa để đo lường sự đa dạng và tránh lặp từ.

    - **Thước đo:** Độ tinh vi của từ vựng (Lexical Sophistication).
    - **Phương pháp:** So sánh các từ được sử dụng với một cơ sở dữ liệu từ vựng chuẩn (ví dụ: CEFR-J, WordNet) để xác định tần suất và "độ khó" của từ vựng. Điều này định lượng hóa việc sử dụng "từ vựng ít phổ biến".
- **Cohesion and Coherence (Sự mạch lạc và Gắn kết):**
    - **Thước đo:** Phân tích các dấu hiệu diễn ngôn (Discourse Marker).
    - **Phương pháp:** Sử dụng các quy tắc hoặc một bộ phân loại để xác định và đếm việc sử dụng các từ/cụm từ chuyển tiếp (ví dụ: "however," "in addition," "as a result").[42] Việc lạm dụng các từ nối đơn giản ("and," "but") có thể bị trừ điểm.
- **Relevance and Completeness (Sự liên quan và Đầy đủ):**
    - **Thước đo:** Độ tương đồng ngữ nghĩa (Semantic Similarity).
    - **Phương pháp:** Đối với Phần 3-5, tạo các vector nhúng câu (sentence embeddings) bằng các mô hình như Sentence-BERT cho cả câu hỏi và câu trả lời của người dùng. Tính toán độ tương đồng cosine giữa chúng để có được một điểm số định lượng cho sự liên quan của chủ đề.[42] Đối với Phần 4, phương pháp này có thể được dùng để kiểm tra xem tất cả các điểm thông tin chính từ văn bản gốc có xuất hiện trong câu trả lời hay không.

Hai mô-đun AI này có mối quan hệ cộng sinh và phải được tích hợp theo một trình tự cụ thể. Đầu ra của Azure Speech (văn bản được chuyển đổi) chính là *đầu vào* cho mô-đun NLP. Một lỗi trong quá trình chuyển đổi giọng nói thành văn bản sẽ lan truyền và gây ra lỗi trong phân tích NLP. Ví dụ, nếu Azure chuyển đổi sai "I think the policy is effective" thành "I think the police is effective", công cụ kiểm tra ngữ pháp NLP có thể báo lỗi ("is" thay vì "are"), và mô hình tương đồng ngữ nghĩa sẽ phát hiện sự lệch chủ đề. Điều này cho thấy sự cần thiết của một điểm số tin cậy (confidence score) từ công cụ speech-to-text. Nếu độ tin cậy thấp, điểm số từ NLP nên được giảm trọng số, hoặc người dùng nên được yêu cầu ghi âm lại.

Bảng 3: Ánh xạ Tiêu chí TOEIC vào Công nghệ Đánh giá AI

Tiêu chí TOEIC	Khía cạnh Đánh giá	Mô-đun Công nghệ	Thước đo / Tính năng API Cụ thể	Ghi chú Triển khai
Pronunciation	Độ chính xác âm vị	Azure Speech	pronunciation Assessment.accuracyScore	Sử dụng chế độ "Reading" cho Phần 1, "Speaking" cho các phần còn lại.
Fluency	Nhịp độ nói &	Azure Speech	pronunciation Assessment.flu	Phân tích các khoảng dừng

	Khoảng lặng		encyScore	không tự nhiên.
Intonation & Stress	Ngữ điệu & Trọng âm	Azure Speech	pronunciation Assessment.prosodyScore	Đánh giá sự tự nhiên của ngữ điệu.
Grammar	Mật độ lỗi & Độ phức tạp	NLP	API kiểm tra ngữ pháp (số lỗi), Phân tích cú pháp (độ sâu cây)	Phạt lỗi, thưởng cho sự phức tạp.
Vocabulary	Đa dạng & Tinh vi	NLP	Tỷ lệ Type-Token, So sánh với danh sách từ CEFR	Thưởng cho từ vựng đa dạng và ít phổ biến.
Relevance	Tương đồng ngữ nghĩa	NLP	Độ tương đồng Cosine của Sentence-BERT	So sánh vector nhúng của câu trả lời với câu hỏi.
Cohesion	Sử dụng từ nối	NLP	Phân tích dấu hiệu diễn ngôn	Đánh giá việc sử dụng hợp lý các từ/cụm từ chuyển tiếp.

Phần 5: Thuật Toán Chấm Điểm Tích Hợp và Cơ Chế Phản Hồi

Phần cuối cùng này đề xuất một mô hình để tổng hợp các điểm dữ liệu từ các mô-đun AI thành một điểm số cuối cùng, đáng tin cậy và cung cấp phản hồi hữu ích cho người học.

5.1. Xây dựng các Thuật toán Chấm điểm có Trọng số cho Từng Phần

Cần phải xây dựng một mô hình để kết hợp các thước đo từ Phần 4 thành một điểm thô duy nhất cho mỗi câu hỏi. Các trọng số phải khác nhau cho từng phần của bài thi, phản ánh tầm quan trọng tương đối của mỗi kỹ năng trong từng nhiệm vụ.

Ví dụ về thuật toán cho Phần 2 (Describe a Picture):
$ĐiểmThô_{P2} = (w_1 \cdot \text{Azure\_Accuracy}) + (w_2 \cdot \text{Azure\_Fluency}) + (w_3 \cdot \text{Azure\_Prosody}) - (w_4 \cdot \text{NLP\_GrammarErrors}) + (w_5 \cdot \text{NLP\_LexicalDiversity}) + (w_6 \cdot \text{NLP\_SyntacticComplexity})$
Trong công thức này, các trọng số ($w_1, w_2, \ldots$) phải được xác định một cách thực nghiệm. Cách tiếp cận tốt nhất là huấn luyện một mô hình hồi quy (regression model) trên một tập dữ liệu các câu trả lời đã được chuyên gia con người chấm điểm. Các thước đo từ AI (Azure_Accuracy, NLP_GrammarErrors, v.v.) sẽ là các biến đầu vào (features), và điểm số của chuyên gia là biến mục tiêu (target variable). Quá trình này cho phép hệ thống "học" được cách các giám khảo con người cân nhắc các yếu tố khác nhau để đưa ra điểm số cuối cùng, thay vì dựa vào các quy tắc heuristic cứng nhắc.

5.2. Hiệu chỉnh việc Chuyển đổi sang Điểm Quy đổi

Vì bảng chuyển đổi chính thức của ETS là bí mật [24], hệ thống cần tạo ra một hàm ánh xạ gần đúng.

- **Phương pháp:** Đề xuất một hàm ánh xạ phi tuyến tính (ví dụ: hàm logistic hoặc một bảng tra cứu) để chuyển đổi tổng điểm thô (tổng điểm của 11 câu hỏi) sang thang điểm 0-200.
- **Hiệu chỉnh:** Hàm này phải được hiệu chỉnh bằng cách sử dụng các mô tả cấp độ năng lực chính thức.[19] Ví dụ, các ngưỡng điểm thô tương ứng với các điểm cắt 130, 160, và 190 trên thang điểm quy đổi phải được xác định thông qua thử nghiệm và phân tích dữ liệu trên một tập hợp lớn các bài thi mẫu.

5.3. Thiết kế Hệ thống Phản hồi Đa tầng, có Tính Hành động

Điểm số cuối cùng chỉ là một phần của giá trị. Tiềm năng học tập thực sự đến từ phản hồi chi

tiết. Một hệ thống phản hồi hiệu quả nên được thiết kế theo nhiều lớp:

- **Lớp 1: Điểm tổng thể và Cấp độ Năng lực.** Hiển thị điểm 0-200 và cấp độ tương ứng (ví dụ: "Level 7: 160-180").
- **Lớp 2: Phân tích theo Tiêu chí.** Hiển thị các điểm thành phần cho Phát âm, Độ trôi chảy, Ngữ pháp, Từ vựng, và Nội dung. Điều này giúp người dùng xác định điểm mạnh và điểm yếu của mình.
- **Lớp 3: Phản hồi Chi tiết, Tích hợp trong Văn bản.** Hiển thị văn bản đã được chuyển đổi của người dùng với các phần được tô sáng tương tác:
    - Các từ bị Azure đánh dấu Mispronunciation có thể được nhấp vào để nghe cách phát âm đúng.
    - Các câu có lỗi ngữ pháp do mô-đun NLP phát hiện có thể được tô sáng kèm theo gợi ý sửa lỗi.
    - Các từ vựng lặp lại hoặc cơ bản có thể được đánh dấu kèm theo gợi ý về các từ đồng nghĩa tinh vi hơn.
    - Các khoảng lặng bị FluencyScore của Azure đánh dấu có thể được hiển thị trên dòng thời gian của âm thanh để chỉ ra nơi độ trôi chảy bị gián đoạn.

Kết luận và Khuyến nghị

Việc xây dựng một hệ thống chấm điểm TOEIC Speaking tự động chính xác là một nhiệm vụ phức tạp, đòi hỏi sự kết hợp sâu sắc giữa kiến thức về khảo thí ngôn ngữ và công nghệ AI tiên tiến. Báo cáo này đã cung cấp một lộ trình toàn diện, từ việc giải cấu trúc bài thi và các tiêu chí chấm điểm đến việc đề xuất một kiến trúc kỹ thuật cụ thể sử dụng Azure Speech và các mô hình NLP.

Các khuyến nghị chính cho việc triển khai dự án:

1. **Ưu tiên Xây dựng Rubric Chi tiết:** Nền tảng của một hệ thống chấm điểm chính xác là một bộ rubric được định nghĩa rõ ràng. Các rubric được đề xuất trong Phần 3, đặc biệt là Bảng 2 và Bảng 3, nên được xem là tài liệu cốt lõi để định hướng cho việc phát triển thuật toán.
2. **Triển khai theo Từng Mô-đun:** Tách biệt việc phát triển Mô-đun 1 (Azure Speech) và Mô-đun 2 (NLP). Điều này cho phép kiểm tra và hiệu chỉnh từng thành phần một cách độc lập trước khi tích hợp chúng.
3. **Tập trung vào Việc Xây dựng Tập dữ liệu:** Thuật toán chấm điểm sẽ chỉ thực sự "chuẩn" khi được huấn luyện và hiệu chỉnh trên một tập dữ liệu lớn gồm các bài nói đã được chuyên gia con người chấm điểm. Chiến lược dài hạn của dự án nên tập trung vào việc thu thập và gán nhãn cho dữ liệu này. Có thể tích hợp một tính năng cho phép người dùng yêu cầu chuyên gia chấm điểm (dưới dạng dịch vụ trả phí) để vừa tạo ra giá trị cho người dùng, vừa xây dựng tài sản dữ liệu quý giá này.

4. **Thiết kế Phản hồi Lấy người học làm trung tâm:** Giá trị lớn nhất của nền tảng không nằm ở việc đưa ra một con số, mà ở việc cung cấp những phản hồi chi tiết, có tính hành động giúp người dùng cải thiện. Hệ thống phản hồi đa tầng được đề xuất trong Phần 5.3 là rất quan trọng để tạo ra một công cụ học tập hiệu quả.
5. **Quản lý Kỳ vọng:** Hệ thống AI, dù tiên tiến, vẫn có những hạn chế và sai số so với giám khảo con người, những người được ETS đào tạo qua một quy trình 10 bước nghiêm ngặt.[8] Do đó, điểm số do hệ thống cung cấp nên được định vị là một công cụ chẩn đoán và ước tính chính xác cao, chứ không phải là một "điểm thi TOEIC chính thức". Sự minh bạch về cách thức hoạt động và những hạn chế của hệ thống sẽ xây dựng lòng tin với người dùng.

Bằng cách tuân theo kế hoạch chi tiết này, dự án có thể tạo ra một nền tảng luyện thi TOEIC Speaking không chỉ mô phỏng chính xác định dạng bài thi mà còn cung cấp một hệ thống chấm điểm và phản hồi tự động, tinh vi, giúp người học chuẩn bị hiệu quả cho kỳ thi thực tế.