研究課題	生成 AI と音声技術を活用した小中高校生向けの診断的な英語スピーキング力テストの開発
副題	～ICT 活用と様々な英語スピーキング活動との親和性の考察～
キーワード	生成 AI／ChatGPT／音声対話／診断的評価／自動採点／手動採点／セルフ・モニター／メタ認知
学校/団体 名	公立全国英語教育研究団体連合会研究部
所在地	〒111-0041　東京都台東区元浅草 1-6-22
ホームページ	

１．研究の背景

GIGA スクール構想の進展により、学校現場では 1 人 1 台端末の常時活用が前提となった。英語科では、リスニングや読解に比べて、スピーキング指導は「授業内での発話量の確保」「個別フィードバックの困難さ」「評価の時間的負担」という課題を抱え続けている。また、スピーキング評価は教師による主観が入りやすく、信頼性（採点者間一致）を担保するには訓練と時間を要する。

一方、近年の生成 AI は音声対話機能、音声認識（自動文字書き起こし）、言語分析（語彙・文法・構成）、フィードバック生成（改善提案）を統合的に提供し得る。もし学習者が自宅等で生成 AI と音声対話を行い、対話内容の文字化と診断的フィードバックを即時に得られるなら、従来の授業内活動に加えて、授業外の発話量増加とセルフ・モニター（自己修正）を促進できる可能性がある。

しかし、生成 AI による自動評価が、経験豊富な教員による手動評価をどの程度反映するか、また、生成 AI を介した学習が学習者のスピーキング力と学習意識（不安、苦手意識、自己効力感、メタ認知）にどのような変化をもたらすかは、学校現場の実践データに基づく検証が必要である。昨年度の研究では、評価基準とプロンプト設計を工夫することで自動評価の実用可能性が示唆された。本年度はそれを発展させ、対話型活動を中核とする診断的指導を実装し、学習成果を実証的に検証した。

２．研究の目的

本研究の目的は以下の四点である。

(1) 生成 AI による自動評価得点と、複数のベテラン教員による手動評価得点の関係を統計的に検証し、自動評価の妥当性・信頼性を明らかにする。

(2) 生成 AI との音声対話、書き起こし、診断的フィードバックを組み合わせた学習（PDCA 型の個別学習）が、学習者の英語スピーキング力（対話・スピーチ・応答）を有意に向上させるかをプリ・ポストテストで検証する。

(3) 学習者のセルフ・モニター力、メタ認知力、学習意欲、英語で話すことへの不安・抵抗感の変化を質問紙調査と自由記述から明らかにし、生成 AI 活用の教育的価値を検討する。

(4) 現プロジェクトの研究者全員が生成 AI のプロンプトを理解・活用し web 課題を作成でき

るようになる。

３．研究の経過　代表的な実践

（1）事前スピーキングテストの実施

本研究では、学習者の英語スピーキング能力およびスピーキングに対する意識の現状を把握することを目的として、9月末に事前スピーキングテストを実施した。対象は中学3年生および高校2年生とし、発達段階の異なる学習者において、本研究で導入する診断的スピーキング学習がどのような影響を及ぼすかを比較検討できるよう設計した。

スピーキングテストは2問構成とし、第1問（Q1）では、日常的な会話場面を想定したロールプレイ形式の会話課題を設定した。これは、実際のコミュニケーション場面に近い状況において、学習者がどの程度自発的に英語を運用できるかを測定することを意図したものである。第 2 問（Q2）では、与えられた話題について3分間の準備時間を設けた上で1分間のスピーチを行わせ（Q2-(1)）、その後、スピーチ内容に基づく 3 問の質疑応答（Q2-(2)）を実施した。

1分間スピーチ後の質疑応答は、第1問をYes/No Question、第2問をTrue or False Question、第 3 問を、学習者の発話内容に関連した自由応答形式の Reference Question とし、段階的に認知的負荷が高まる構成とした。これにより、単なる暗記や定型表現の再生ではなく、内容理解、即時的な判断、さらには自分の発話内容を踏まえた応答力を総合的に評価できるよう配慮した。

スピーキングテスト終了後には、学習者のスピーキングに対する態度や意識の変化を測定するため、アンケート調査を実施した。アンケートでは、「スピーキングを行う前」「スピーキングを行っている最中」「スピーキングを行った後」の各段階において、学習者が意識している点や心がけている行動について、「6. かなりそう思う」から「1. 全くそう思わない」までの6段階で回答させた。これにより、スピーキングへの不安感、自己効力感、振り返り意識などを多面的に把握することを試みた。

提出物としては、①発話内容を自動文字起こししたテキストを貼り付けた Word ファイル、②スマートフォンで録音した音声ファイル、③ChatGPT による自動採点結果を貼り付けた Excel ファイルの 3 点を求めた。ChatGPT による採点項目は、①発音・アクセント、②流暢さ、③語彙、④文法、⑤一貫性、⑥情報量、⑦総合評価の7項目とし、各項目 0 点から 3 点までの 4 段階評価とした。また、評価項目ごとに改善につながる具体的なフィードバックを自動提示する設定とし、学習者が自らの弱点を認識しやすくなるよう工夫した。

最終的に、本事前スピーキングテストには、中学3年生41名、高校2年生19名の計60名が参加した。

（2）診断的スピーキング学習サイクル

事前スピーキングテスト終了後、9 月末から 1 月初めまでの約 15 週間にわたり、ChatGPT を活用した診断的スピーキング学習を継続的に実施した。本実践は、学習者が自宅で自主的に取り組むことを前提とし、教室外学習におけるスピーキング練習の質的向上を目指したものである。

学習者には記録シートを配布し、毎週、自分が取り組む予定の課題に〇を付け、実際に実施した課題にも〇を付けた上で、簡単な振り返りコメントを記入させた。さらに、その内容を週次ア

ンケートとしてオンラインで提出させ、教員が進捗状況や学習傾向を把握できるようにした。

課題内容は、①中高生が日常生活で遭遇する可能性の高い場面設定による会話課題、②1 分間スピーチとそれに続く 3 問の Q&A 課題、③その日に起こった出来事を日記風に話す課題、④自分の興味・関心に基づいて自由に話す課題、の 4 種類とした。毎週、会話場面やスピーチトピックは教員側で変更したが、どの課題を選択するか、また学習計画をどのように立てて実行するかについては、学習者自身に委ねた。

このような設計は、学習者の主体性を尊重すると同時に、自らの学習過程を振り返るメタ認知能力の育成を意図したものである。そのため、教員側からは基本的に個別指示や助言は行わず、データ管理および全体傾向の把握に専念した。

学習者は、①音声対話による発話、②自動文字起こしによる発話内容の可視化、③語彙・文法・内容・流暢さ等に関する診断的フィードバックの取得、④次回に向けた振り返り、という一連の PDCA サイクルを繰り返した。この過程において、多くの学習者が「自分では正しく言えたつもりであった表現が、文字起こしでは別の語として認識されていた」ことに気づき、発音や音の連結、アクセントといった要素への意識を高めていった。

また、自分の言いたい内容を表現しようとする中で、語彙や文法知識の不足を実感し、それを次回の課題として自ら設定する学習行動が見られるようになった。これらの点から、本実践は単なる発話量の増加にとどまらず、自己修正（セルフ・モニタリング）を伴う質的に高いスピーキング学習を成立させたと考えられる。

さらに教員側にとっても、記録シートや提出ログを通じて学習者の実施量や共通課題を把握できたことで、授業内において「聞き返し方」「言い換え表現」「発話の間の取り方」など、短時間で効果的に補強すべき指導内容を明確にすることが可能となった。

（3）事後スピーキングテストの実施

2 月中旬に、事後スピーキングテストとして、事前スピーキングテストと同一構成の 2 問からなるテストを実施した。会話場面およびスピーチトピックについては、事前テストと同程度の難易度となるよう配慮しつつ、異なる内容を設定した。

また、スピーキングに対する意識・態度を測定するアンケートについても、事前テストと同一の項目および評価尺度を用いて実施し、学習前後の変化を比較できるようにした。

（4）採点・分析

2 月中旬から 3 月初めにかけて、事前・事後スピーキングテストの採点およびアンケート結果の分析を行った。

スピーキングテストの採点は、中学校および高等学校の指導経験が豊富な教員複数名が、事前・事後それぞれの音声を個別に評価した後、協議を通じて最終的な点数を確定させる方法を採用した。評価項目は ChatGPT の自動採点と同一とし、その結果を比較することで、自動採点の妥当性についても検証した。

アンケートについては、6 段階評価の数値変化を定量的に分析するとともに、自由記述の内容

から、学習者の意欲、メタ認知、セルフ・モニター力の変容を質的に分析した。

以上のように、本研究では、事前・事後のスピーキング・テストおよび15週間にわたる診断的スピーキング学習サイクルを通して、学習者のスピーキング能力の変化、スピーキングに対する態度やメタ認知の変容、ならびにChatGPTによる自動評価の妥当性について検討し、本研究の成果と教育的示唆について多面的に捉える実践を行い、考察した。

４．研究の成果・分析

(1) スピーキング力の変化：

生成AIとの毎週の音声対話学習（PDCA型の個別学習）を15週間続けたことが、学習者の英語スピーキング力（対話・スピーチ・応答）を有意に向上させたかを検証するため、プリテストの音声の手動評価とポストテストの音声の手動評価の値を比較した。課題を継続的に実施した学習者の 62.5%以上で総合得点が向上し，スピーキング力の伸長が確認された。つまり，今回の調査では，スピーキング力を対話力，スピーチ力，スピーチ内容の質疑応答の3種類のテストを実施した。この3つの中で，ポスト・テストの平均点がプリ・テストの平均点より上昇したのは，スピーチ力(+0.5点）と質疑応答力（+0.63点）であった。一方，対話力については，むしろポストテストの平均点の方が低く（-0.25）なる現象が見られた。この結果は，生成AIと学習者が音声対話をすることは，練習を積み重ねるだけでは難しく，何らかの指導を行う必要性を示していると思われる。一方，テーマについて1分間のスピーチを行い，学習者自身のスピーチに対するAIの質問に足して，学習者が解答するという質疑応答するスピーキング力は，練習を積み重ねることによって，伸びていくことを示していると考えられる，なお，これらの平均点の差は統計的に有意ではなかった。有意な結果にならなかった原因として，学習者の音声がきちんと録音されていないものがあり，有効な対象人数が24名と少なかったことが考えられる。今度，対象人数を増やして再度調査してみる必要性があると思われる。

(2) 課題実施量（週ごとの実施回数）と得点伸長の関係

プロジェクトに取り組んだ中学生は15週なので，課題に取り組んだ量をABCに分けて，課題に取り組んだ量とスピーキング力の伸びの関係を分析した。A は高頻度（30 回以上、週２以上），B は中頻度（30回未満15回以上、週２未満週１以上）C は低頻度（14回以下、週１以下）の回数で分けた。スピーキング・テストの得点の伸びをポスト・テストの平均得点からプリ・テストの平均得点として計算した。その結果，高頻度のグループ A は-0.1 となり伸びは見られなかった。一方，中程度のBグループは0.31点伸び，低頻度のCグループは0.57点の伸びが見られた。ただし，プリテストとポストテストの得点差は，繰り返しのあるｔ検定の結果，有意な結果にはならなかった(p=0.737)。

当初，課題に取り組んだ量が多いほど，スピーキング・テストの平均点は伸びると予測していたので，この予測とは反対の結果となった。この理由として，課題に取り組んだ量が多いグループの学習者は全体として英語力が高く動機づけも高い生徒が多かった。このため，もともと高いスピーキング力をもっていたため，プレ・テストとポスト・テストの間で差が出なかったと考

えられる。これはHarpe（2015）が提唱する天井効果の1つである可能性がある。一方，課題に取り組んだ量が少ないBやCのグループは，少しの課題量でも，今まで慣れていなかった英語スピーキングに慣れてきて，プレ・テストとポスト・テストの得点差がはっきりと現れた可能性がある。なお，この傾向は高校生のプリテストとポストテストの得点差と取り組んだ課題量についても同様の傾向が見られた。

（3）生成AIによる自動評価の有効性の検証：AI評価と手動評価の相関

対話力，スピーチ力，スピーチ内容の質疑応答の3種類のテストについて，AI評価と手動評価をスピアマンの順位相関係数（ρ）を計算したところ，表1のような結果となった。このことから，対話力については弱い相関，スピーチ力については中程度の相関が見られた。このことから，さらに改良を加えれば将来的に生成AIによる評価は一定の妥当性および信頼性を有することが示唆された。このことから、生成AIによる自動評価は、教師による手動評価を補完する手段として活用できる可能であると考えられる。しかし，質疑応答力についてはほとんど相関が見られなかった。質疑応答力については，改めて評価規準（ルーブリック）を再検討する必要があると思われる。

表1

区分	相関係数	p値
対話力	ρ =0.40	p=0.15
スピーチ力	ρ =0.47	p=0.06
質疑応答力	ρ =0.09	p=0.72

（4）メタ認知能力・学習意欲・セルフモニター力への影響についての検証：

本研究で実施したアンケート項目（プレとポスト）のうち、生徒のメタ認知能力・学習意欲・セルフモニター力に特に関係が深いと考えられる項目を整理した。（ ）内の数値は、6点満点中の平均値の変化量を示している。

1.メタ認知能力に関しては、「自分の学習・発話を客観視し、計画・調整・振り返りができているか」を測るということに関係する項目を集めた。「話す前に何を言うかを考えてから話すようにしている」(4.9→5.0,+0.1)「自分にとって話しやすい方法（ゆっくり話すなど）を考えて準備する」(4.4→4.8,+0.4)「うまく伝えられているかを考えながら話す」「話した後に、どこがよかったか、うまくいかなかったかをふり返って考える」(4.3→4.6,+0.3)「もっとよい言い方があったかどうかを考える」(4.7→4.8,+0.1)「英語を話す力が少しずつ上達しているか気にしている」(4.9→4.9,+0）といった項目が該当する。これらは、生徒が自らの発話を事前・最中・事後の各段階で客観的に捉え、調整しようとする意識を測定するものであり、メタ認知的方略の使用を示す指標であると考えられる。一つの項目を除く全ての項目において数値がプラスになっていることから、メタ認知能力が上がったことが伺える。

2.学習意欲に関しては、「英語を話すのは楽しいと感じる」(3.0→3.2,+0.2)「英語で話せると、またがんばろうと思う」(4.3→4.4,+0.1)「苦手でも英語を話すことに意味があると思う」(4.9→5.0,+0.1)「英語を話すことは将来自分にとって何らかの役に立つと思う」(5.5→5.3,-0.2)の項目が該当する。これらの項目は、英語で話す活動に対する肯定的感情や価値認知を反映しており、生徒の内発的動機づけや学習継続意欲を把握する上で重要である。一つの項目を除く全ての項目において数値がプラスになっていることから、学習意欲が上がったことが伺える。

3.セルフモニター力に関しては、「言いたいことがうまく言えなかったときに、別の言い方を考える」(4.9→5.1,+0.2)「話していて言葉に詰まったら、話がとぎれないように工夫する」(4.8→5.1,+0.3)「英語を話すとき、落ち着いて話そうと意識している」(4.1→4.6,+0.5)「うまく伝えられているかを考えながら話す」(4.6→4.9,+0.3) という項目が該当する。これらは、発話中に自らの言語産出を観察し、必要に応じて修正・補完を行う力を測るものであり、実際のスピーキング場面における即時的な自己調整能力を示していると考えられる。全ての項目で、数値がプラスとなっており、セルフモニター力が上がったことが伺える。

以上のことから、本アンケートは、生徒のスピーキング活動における認知的側面と情意的側面の双方を捉える構成となっており、メタ認知能力・学習意欲・セルフモニター力の把握に一定の妥当性を有すると考えられ、その数値は有意に変化していることから、今回の研究で実施したスピーキング活動の生徒の情意面へのプラスの影響が確認されたといってよい。

(5)生徒の自由記述コメントに見られる3要素の関係性：

プレとポストの際のアンケート、週ごとに取った生徒のアンケートのコメント欄からうかがえることを以下のように分析した。

自由記述の分析からは、上記3要素が相互に関連しながら発現している様子が確認された。例えば、「質問に答えるときに、授業ではよく考えてから答えられるが、AIとの会話ではそうではなく、自分が文法や表現を十分に理解していないことに気づいた」という記述からは、学習環境の違いを通して自らの理解度や課題を客観的に捉えるメタ認知的気づきが読み取れる。また、「聞き取れなかったり、間違えた答えを言ってしまったが、もう一度聞こうと思った」という記述には、発話や理解の不十分さを認識した上で再挑戦しようとする姿勢が示されており、セルフモニター力と学習意欲の両面が表れている。さらに、「新しく習った文法や単語を使うことを意識して話したい」「単語をしっかり学び、文法を定着させたい」といった記述からは、自身の課題を把握した上で次の学習行動を具体的に構想する様子が見られ、メタ認知能力と学習意欲の結びつきが確認できる。中でも、「自分の考えを、簡単な英語でいいから声に出してみることが大切だと思った」という記述は、自己の言語レベルを踏まえた発話方略の選択（メタ認知)、発話中の調整意識（セルフモニター力)、および継続的に取り組もうとする姿勢（学習意欲）が同時に表れており、3要素が統合的に機能している例として注目される。

これらの自由記述から、AI を活用したスピーキング活動は、生徒に自らの理解度や発話の在り方を振り返らせる契機となり、メタ認知能力・学習意欲・セルフモニター力を相互に高め合う可能性を有していることが示唆された。

以上の結果から、本研究で実施したスピーキング活動およびアンケート調査は、生徒の英語運用能力そのものだけでなく、学習を調整・持続させるための内的要因に働きかける効果を有していると考えられる。特に、メタ認知能力・学習意欲・セルフモニター力が相互に関連しながら育成されている点は、本研究の重要な成果である。

５．今後の課題・展望

本研究は、生徒の ICT 環境や各自治体の運用方針に大きく左右され、多数の学校が同一条件で参加することには一定の制約があった。また、生成 AI の応答挙動を制御するためにプロンプトの継続的な改善を行ったものの、生徒からは、発話中の相槌や指示、質問提示のタイミング、採点への移行に関する課題が指摘された。今後は、より安定した動作を実現するためのプロンプト設計およびシステム面での改善が必要である。

さらに、本プロジェクトでは、研究者全員が生成 AI を活用した課題作成を行える体制構築を目標としたが、学校ごとの端末環境や利用制限の違いにより、十分な共有が難しかった点も課題として残った。

６．おわりに

本研究は、英語教育学・教育工学の専門家である創価大学の山内豊教授、そして生成 AI（ChatGPT）と音声工学を専門とする東京大学の峯松信明教授のご尽力なくしては成し得ませんでした。お二方には深く感謝申し上げます。また、スピーキングテストの作成、音声評価の相関の統計処理等に協力してくださった創価大学の沖谷瑞保さんにも謝辞を申し上げます。さらに、多忙な学校業務の合間を縫い、週末も含め頻繁に会議に参加し、議論を重ねてくださった全英連研究部会の先生方に感謝の意を表します。最後に、音声録音、スピーキング・テストの実施、英語スピーキング訓練等に快く参加してくださった生徒たちに心より感謝申し上げます。

７．参考文献

Harpe, S. E. (2015). How to analyze Likert and other rating scale data. Currents in Pharmacy Teaching and Learning, 7(6), 836-850.

Tekin, G., & Kaya, S. (2026).Emotional intelligence and metacognitive awareness in predicting foreign language speaking self-efficacy. *Frontiers in Psychology, 17,* 1721694. https://doi.org/10.3389/fpsyg.2026.1721694

Yamauchi, Y. & Nishikawa, M. (2023). Which type of speaking test predicts L2 overall proficiency most? Proceedings of Asia TEFL 2023.

（その他、関連研究・研究会資料等）