Using automatic speech recognition technology to enhance EFL learners' oral language complexity in a flipped classroom

Michael Yi-Chao Jiang, Morris Siu-Yung Jong, Wilfred Wing-Fat Lau, Ching-Sing Chai
Department of Curriculum & Instruction and Centre for Learning Sciences & Technologies, The Chinese University of Hong Kong, Hong Kong S.A.R, China

Na Wu
Department of Foreign Languages, Shenyang Normal University, Shenyang, China

The present study examined the effects of using automatic speech recognition (ASR) technology on oral complexity in a flipped English as a Foreign Language (EFL) course. A total of 160 undergraduates were enrolled in a 14-week quasi-experiment. The experimental group (EG) and the control group (CG) were taught with a flipped approach, but the EG students needed to undertake an additional pre-class task with ASR technology. In each unit, all students' in-class task performance was recorded, based on which the metrics of oral complexity were coded and computed. A two-way between- and within-subjects repeated measures design was conducted to examine the effects of the group factor, the time factor and the group × time interaction effects. The results showed that the EG students performed statistically better than their counterparts in the CG on lexical complexity and syntactic complexity. Moreover, significant improvement in phrasal complexity was witnessed over time in both groups. Significant group × time interaction effects were witnessed on overall complexity or subordination complexity. The gradients of the EG trajectories of the two metrics were greater than those of the CG. However, on phrasal complexity, the interaction effect was not significant.

Implications for practice or policy:

- EFL teachers could integrate ASR technology into pre-class tasks to improve students' oral English complexity.
- EFL teachers need to be aware that phrasal complexity may be more sensitive to flipped EFL instruction than overall complexity and subordination complexity.
- Course developers could integrate ASR technology in fostering EFL learners' overall complexity and subordination complexity.

Keywords: flipped classroom approach (FCA), automatic speech recognition, oral complexity, English as a Foreign Language (EFL)

Introduction

The flipped classroom approach (FCA) has gained tremendous momentum in language education in the past few years. Many empirical studies have been conducted to prove that the FCA is more effective than a regular lecture-based one (Blair et al., 2016; Fox & Docherty, 2019; Jong et al., 2019; Lee & Wallace, 2018; Ozudogru & Aksu, 2020). Over the past decade, numerous language teachers and researchers have tried various methods to flip their instruction to examine the effectiveness of the FCA. However, some synthesis studies on various flipped courses pinpointed that students may have difficulty preparing themselves adequately before class, which may result in a low level of preparedness (Akçayır & Akçayır, 2018; H. C. Lin & Hwang, 2019).

In light of the technologies integrated into flipped language classrooms, most courses involve only some easy-to-use technologies such as online videos and exercises (Song et al., 2017). In contrast, cutting-edge technologies that could help establish an immersive learning environment are under-leveraged (Chien et al., 2020; Geng et al., 2019; M. Y. C. Jiang et al., 2020; H. C. S. Lin et al., 2019). It is advisable for language teachers to enhance students' preparedness through integrating technologies such as virtual reality and automatic speech recognition (ASR) into pre-class self-learning. This may enhance students' oral competencies especially in learning foreign languages.

In foreign or second language (FL or L2) research, complexity is investigated as a basic descriptor of target language performance and as an indicator of target language proficiency (Bulté & Housen, 2012). Although many existing studies have obtained empirical evidence in favour of the effectiveness of the FCA with FL or L2 instruction, the outcome variables investigated were mostly students' overall language performance or general language proficiency. Since target language proficiency is no longer perceived as a unitary construct but multi-componential, it is necessary for FL or L2 researchers to examine the effects of the FCA on learners' linguistic performance from more concrete angles (N. Ellis & Robinson, 2008; Norris & Ortega, 2009). However, few studies have adopted domain-specific indicators (e.g., FL or L2-related measures of complexity, fluency and accuracy) to examine the effectiveness of the FCA. Studies with more refined indicators of learners' linguistic performance can contribute to a deeper understanding of the FCA for language learning and diversify the instructional design of the FCA.

Related works

Learner preparedness in flipped English as a Foreign Language (EFL) classrooms

As proposed by Thorndike (1932), one law of learning is students' readiness to learn, which can strongly influence the degree of success achieved (see also Jong et al., 2006, 2013). The FCA inverts the time and place where lectures and homework should occur. Typically, knowledge delivery through computer-mediated lectures is flipped outside of class while higher-order knowledge construction tasks are the focus for in-class activities with instructor and peer support (Bergmann & Sams, 2012; Jong, 2017, 2019; Jong et al., 2019). In flipped learning, students are supposed to prepare themselves with the pre-class content to achieve a proper level of readiness. Such readiness is referred to as learner preparedness in the present study, emphasising learners' active and targeted preparation for in-class learning and activities. Students' preparedness in flipped classrooms may directly influence their engagement in in-class activities and their academic achievement (Rahman et al., 2015; Stockwell, 2008; Sun & Xie, 2020). However, few empirical studies have attempted to investigate the effects of learner preparedness on students' in-class performance (Sun & Xie, 2020).

Although learner preparedness is an identified necessary condition for the success of the FCA, some review studies have found that most flipped courses focused on essential knowledge and skills training (H. C. Lin & Hwang, 2019) and teachers had practical difficulty in developing students' higher-order thinking skills (Jong, 2015; Lee & Wallace, 2018). In a flipped EFL setting, the possible discrepancy between what is flipped outside of class and students' acquired skills in the target language may lead to inadequate learner preparedness for in-class activities, especially those involving oral English. Thus, the flipped classroom may fail to repurpose the in-class time and may become less structured. In response, technology-aided mediating tasks may help bridge the gap.

According to M. Y. C. Jiang et al.'s (2020) synthesis work, educational technology is not fully harnessed in flipped language classrooms. Even though watching YouTube videos and doing online quizzes for self-study was the most common way of using technology in flipped courses, some more advanced technologies involving artificial intelligence (e.g., ASR technology) are yet to be widely utilised. The FCA should go beyond YouTube videos plus drills and need further innovation. In particular, some researchers have pointed out that there is little understanding of the role and significance of ASR technology in computer-assisted language learning (Golonka et al., 2012; Steel & Levy, 2013).

Integrating ASR technology into FL learning

Information and communications technology can provide language learners with opportunities to receive enhanced input, engage in interaction and improve linguistic production (Chapelle, 2003; Lan et al., 2018). ASR technology can be harnessed to help students with their oral proficiency and arguably hold forth the promise of supporting different types of oral practice and providing real-time feedback on many aspects of language proficiency, including pronunciation quality and target language use (Franco et al., 2010). ASR-based applications such as computer-assisted pronunciation training can provide benefits that cannot be easily achieved in traditional classrooms (Evers & Chen, 2020), including significant amounts of practice, consistency, the unbiased nature of feedback and diverse forms of visual representations (Levis, 2007).

The most significant advantage of ASR technology is that it can free the language teacher from the massive, one-to-many work of providing frequent feedback for students' drill-and-practice activities. Instead, students can interact with the ASR-based software without time and space constraints until their utterances can be recognised with few errors. Evidently, a teacher can hardly afford enough time to provide individual feedback on every student's FL or L2 expressions in regular language classrooms. Moreover, on the learners' side, oral practice usually needs a substantial amount of time and requires frequent corrective feedback from a source other than the perception of a language learner, making ASR technology "a suitable arena for a tireless computer" (Franco et al., 2010, p. 402). This form of technological support is especially important for FL or L2 learners who suffer a high level of language learning anxiety, which is a major problem identified among EFL learners (e.g., Horwitz, 2002; Y. Jiang & Dewaele, 2020; Nakazawa, 2012).

ASR-based software is a useful means for FL or L2 learners' pronunciation practice that can help them detect frequent errors, enhance their target language pronunciation and make their interlanguage comprehensible (e.g., McCrocklin, 2016). The real-time transcriptions provided by ASR-based software can serve as feedback to evaluate whether their speech is acceptable when learning pronunciation in the target language or performing oral tasks. If their utterances are not recognised or mismatch the transcriptions, learners may need to correct what is articulated. Therefore, ASR-based practice can facilitate the process of proceduralisation of the learners' interlanguage (DeKeyser, 2001, 2007; Segalowitz, 2010), which may result in a degree of automatisation, that is, "a process of development from conscious, controlled and often slow processing of declarative knowledge to more rapid, effortless and attention-free processing of language, in their performance" (Tavakoli et al., 2016, pp. 463–464). Consequently, such acquired proceduralisation may spare learners more room for producing more complex utterances in the target language.

Complexity in EFL research

Complexity is a prominent element of the widely acknowledged complexity, accuracy and fluency (CAF) framework (Skehan, 1996, 1998) in language education. Many FL or L2 practitioners and researchers hold that target language proficiency is a multifaceted rather than a unitary construct and that its principal components can be validly captured by the notions of complexity, fluency and accuracy. In general, the CAF framework defines language proficiency as the complex interplay of the three elements (Norris & Ortega, 2009; Pallotti, 2009; Skehan, 2009b; Tavakoli, 2016). These authentic indicators for the assessment of language performance can be integrated with ASR technology to promote language learning, especially for fostering oral competencies.

Complexity is generally defined as the competence to use a wide and varied range of sophisticated structures and vocabulary in the target language (R. Ellis, 2003, 2008; Housen et al., 2012; Skehan, 1998). Studies using complexity as a dependent variable have produced mixed and sometimes contradictory results (cf. Robinson, 2007; Skehan, 2009b; Spada & Tomita, 2010), which, according to Bulté and Housen (2012), may partially result from the inconsistent definitions and operationalisations of complexity. To be specific, the present study examined linguistic complexity, which, following Bachman (2005), we specified as lexical complexity and syntactic complexity (see also Bulté et al., 2008, for a review on linguistic complexity).

Lexical complexity is an essential indicator of how difficult to read and complex a text is. One of the most applied indicators for assessing lexical complexity is type/token ratio (TTR; Vermeer, 2000). Nonetheless, quantitative linguistic studies have shown that TTR is easily affected by the length of the text sample. The computer-generated vocd-D value can compensate for such disadvantage and thus improve the measurement of lexical complexity (see McCarthy & Jarvis, 2007, for a review on voc-D). It is one of the most reliable measures in the literature (McKee et al., 2000). The vocd-D value is a result of a series of random text samplings and thus needs to be calculated by computer algorithms. Unlike TTR and its derivatives, the vocd-D value is sample-length free and can involve all the words produced by the interlocutors (Albert, 2011). Therefore, it is reliable and consistent (McCarthy & Jarvis, 2007, 2010; McKee et al., 2000). McCarthy and Jarvis (2010) concluded that by using more complicated measures, such as vocd-D, researchers could get a clearer idea of the text as a whole and avoid drawing false conclusions.

The other type of complexity is syntactic complexity and there are both theoretical and empirical justifications for the claim that syntactic complexity must be measured multidimensionally (Norris &

Ortega, 2009). One class of the most widely used metrics across language-related fields is based on the length and is calculated by dividing words by a chosen production unit. Length-based measures are prevalent and commonly employed in developmental language acquisition and the analysis of both written and oral language in linguistic studies. For example, overall complexity or phrasal complexity fall into this category. In contrast, subordination complexity is another measure of syntactic complexity. It is generated by computing the amount of subordination by counting all clauses and dividing them over a chosen production unit (Norris & Ortega, 2009). Analysis of speech unit (AS-unit) is revised based on the extant production units currently available to provide an agreement on the nature of the unit for segmenting problematic oral data (Foster et al., 2000). An AS-unit is "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s)" (Foster et al., 2000, p. 365). They proposed it to deal with the fragmentary nature of oral data and provide a solution against the fuzziness and complexness of the spoken language (such as false starts, self-corrections, interruptions). Therefore, the present study adopted the AS-unit in computing the metrics of complexity.

Formulating research questions

Tseng et al. (2020) suggested that course instructors should raise consciousness in integrating emerging technologies such as ASR technology into their pedagogical practice. However, as mentioned, few studies have integrated ASR technology into flipped EFL instruction. Moreover, FL or L2-specific measures such as language complexity have been barely investigated to explore the effects of the FCA on EFL learners' linguistic performance. Therefore, we formulated the following research questions in response to the research gaps identified:

(1) Do students that prepare themselves with ASR technology outperform their counterparts in the CG in terms of oral English complexity in a flipped EFL setting?
(2) Does students' oral English proficiency in flipped EFL classrooms change significantly over time?
(3) Is there any interaction effect between group and time on students' oral English proficiency?

Methods

Participants and research context

A total of 160 first-year students from a public university in China participated in this study. The study was approved by the university, and the students gave their consent as participants. The average age was 18.1 years old; 21.3% of them were male, and 78.7% were female. A pre-intervention survey was conducted to collect background information about the participants' EFL learning experience. The participants had studied English for an average of 10.9 years before entering university. Due to the examination-oriented learning in their secondary education, 71% of the students reported little experience in oral English learning and practice because English reading and writing skills were predominantly tested in their college entrance examinations. According to the course teacher, most students had difficulties in communicating in oral English. Additionally, 93% of them had little experience in flipped learning.

The participants came from four parallel classes and all registered for the compulsory EFL course College English, which consists of different streams (e.g., integrated, listening and speaking, fast reading). The present course was in an integrated stream, implemented in the fall 2019 semester. The 14-week course aims to develop students' integrated English skills, enhance their ability to use the target language for general and academic purposes and understand and appreciate the culture in English-speaking countries. The students had two 45-minute face-to-face sessions each week. The four tracks were taught all in a flipped fashion with the same online learning platform – Unipus (Figure 1). The students had their sessions on different weekdays by the same course teacher, who has been teaching on the integrated stream for the 10th consecutive year.

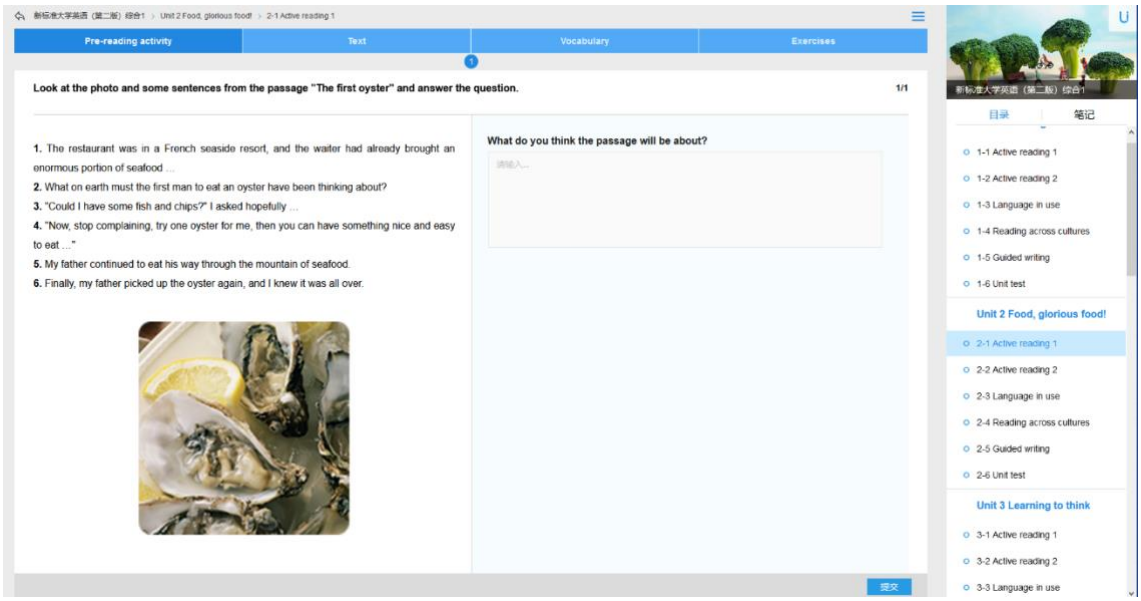


Figure 1. Screenshot of the learning platform Unipus

Instructional design

Before the course began, the four tracks had been randomly assigned into an EG and a CG. The EG students were assigned a mediating ASR-based oral task in addition to the self-learning resources on Unipus for pre-class preparation. In contrast, the CG students were only given the materials on Unipus before class. All the students were randomly assigned to workgroups of two to four for in-class activities within each track.

Figure 2 illustrates the entire instructional process. A placement test was administered in conjunction with the pre-intervention survey before the commencement of the course. Week 1 of the course was scheduled to orient the students to the novel flipped learning method and the course basics, including course assessment and the supplementary content on Unipus. According to Moranski and Henery (2017), orienting students to (re)mediate their understanding of the FCA helps to ensure the success of implementing this new learning approach.

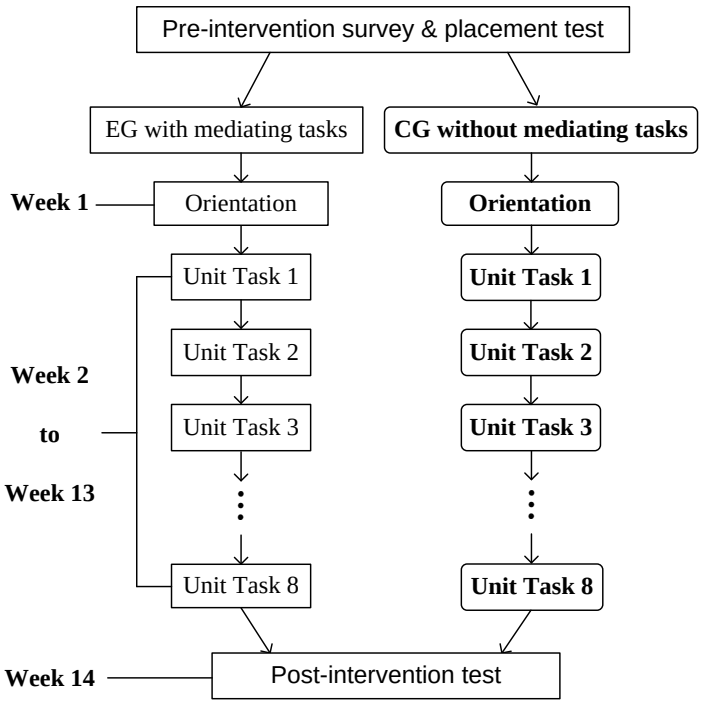


Figure 2. The instructional procedure of the study

The students on each track received self-learning materials that the course teacher decided to flip outside of the regular classroom through Unipus. The materials were developed by the textbook publisher in the form of a massive open online course and could be accessed conveniently by the students with mobile devices or desktop computers. The students received the materials 1 week before each face-to-face meeting; they were required to self-study and raise questions via WeChat (a social media platform) or Unipus.

In particular, for the EG students, the mediating ASR-based tasks facilitated the students' pre-class self-learning orally. These tasks needed to be completed through iFlyRec (Xunfei Tingjian), an ASR-based mobile device application (https://www.iflyrec.com). It realises real-time transcription and translation in multiple languages (e.g., Chinese, English, Korean) and some typical Chinese dialects (see Figure 3 for the user interface of the application). While practising, the students can see how their utterance is "understood" by the application via the transcribed text and spot their pronunciation or morphosyntactic errors in a real-time manner. Besides, it can also translate from English to Chinese or Chinese to English. With this function, when they do not know how to express themselves in English precisely, the students can use Chinese and then ask the application to translate for them and they then learn from the translated expression. This function distinguishes iFlyRec from other ASR-based software.

Figure 3. Screenshots of the iFlyRec application

The EG students needed to complete the mediating tasks by answering orally some follow-up questions based on what they self-studied before each session. While giving their answers, the students were required to use the application to transcribe what they articulated. Using the transcription as real-time feedback, the students could monitor their pronunciation and expression and thus correct themselves where necessary. By self-correcting the errors with the ASR-based feedback, the students can familiarise themselves with the in-class activities in terms of oral expressions and improve their oral performance.

Week 2 to Week 13 covered the instructional practice of eight units. For each unit, the students in both groups undertook the communicative tasks in English in class (see the Appendix for a sample task). Within each workgroup, the students orally expressed their opinions or experiences regarding the unit topic. According to Bloom's taxonomy (Anderson & Krathwohl, 2001), the in-class tasks were designed to elicit the students' authentic language use for the applying-, analysing- and evaluating-oriented peer interaction. While taking turns to speak, the students were required to record their discussions for data collection purposes. The recordings of Units 2, 4, 6 and 8 were used for data analysis, but the students did not know which unit would be analysed. For Week 14, the students in both groups were required to take a post-intervention exam as a summative assessment of the course.

Measures

Following the suggestions by Read (2000) and Skehan (2003), we operationalised lexical complexity as the vocd-D value, which was calculated automatically through computer algorithms (Table 1). We employed the website TextInspector (https://textinspector.com) to estimate the lexical complexity of the students' oral language output. The website has been employed in published works (e.g., Bax et al., 2019).

Table 1
Metrics of oral language complexity

Dimensions and subdimensions		Metrics
lexical complexity		vocd-D value
syntactic complexity	overall complexity	mean length of AS-unit
	phrasal complexity	mean length of clause
	subordination complexity	mean number of clauses per AS-unit

On the other hand, in accordance to the synthesis study by Norris and Ortega (2009), we operationalised syntactic complexity by overall complexity, phrasal complexity and subordination complexity in the present study (Table 1). In general, the three measures are all length-based metrics with a multi-clausal unit of production (i.e., AS-unit) in the denominator. Precisely, the overall complexity measures the general syntactic complexity of the utterance produced by the interlocutors. Thus, we operationalised it as the mean length of an AS-unit. In contrast, phrasal complexity and subordination complexity measure syntactic complexity at the subclausal level. Phrasal complexity measures complexity via phrasal elaboration; thus, we defined it as the mean length of a clause articulated by the interlocutor. Subordination complexity measures the complexity via subordination; thus, we defined it as the mean number of clauses per AS-unit.

Data collection and analysis

After collecting the students' classroom recording data (n = 160), we sorted out, numbered and tested the data for transcription. Due to classroom noise and dropout issues, four workgroups of 14 participants were entirely removed from the transcription. Furthermore, as 18 participants were absent for one or more times during the semester, their data was also removed. Therefore, we eventually adopted and analysed the data transcribed and coded from a total of 128 participants (68 EG students and 60 CG students).

To annotate linguistic performance (e.g., complexity, accuracy and fluency) from audio or video data, we adopted ELAN (https://tla.mpi.nl/tools/tla-tools/elan; see Figure 4 for a screenshot), which enables users to add an unlimited number of textual annotations to audio and videorecordings (Lausberg, & Sloetjes, 2009). Using ELAN, annotations to transcriptions can be created on multiple layers, also known as tiers. These layers can be hierarchically interconnected. An annotation can either be time-aligned to the media or can refer to other existing annotations.

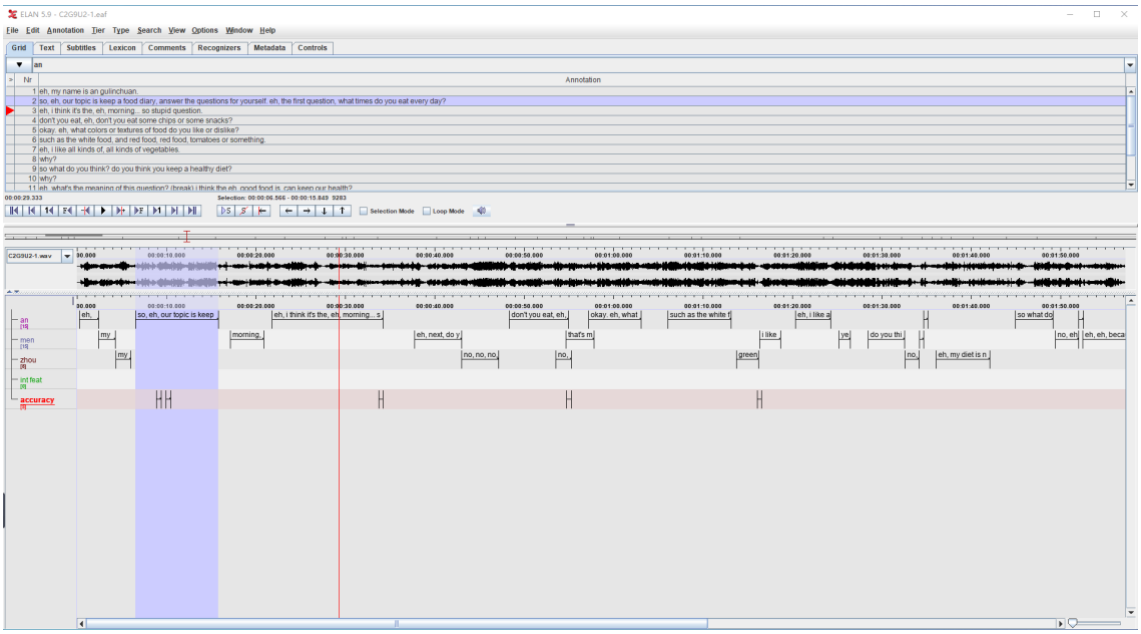


Figure 4. Screenshot of the ELAN workspace

The transcribed recordings of in-class peer interaction were coded into frequencies and relative frequencies (against AS-unit) to form study-generated quantitative data. In response to the research questions, a two-

way repeated-measures ANCOVA (pre-intervention English proficiency controlled for as a covariate), that is, a mixed within- and between-subjects design, was conducted. The independent variables were the group factor (two levels: EG and CG) and the time factor (four levels: Times 1, 2, 3 and 4). The dependent variables were the metrics of oral complexity coded from students' in-class task performance. SPSS and Excel were also employed to store and manage the data.

Findings

Between-subjects effects of ASR technology on students' oral English complexity

Lexical complexity

Table 2 shows the descriptive statistics of the lexical complexity in each group at each time point. The estimated marginal means of the vocd-D value over time was 41.791 (*SE* = 1.971) for the EG students and 36.417 (*SE* = 1.851) for the CG students. For the time factor, the estimated marginal means of vocd-D across the two groups was 35.762 for Time 1 (*SE* = 2.452), 36.371 (*SE* = 2.044) for Time 2, 39.837 (*SE* = 1.298) for Time 3 and 44.445 (*SE* = 1.322) for Time 4, showing a clear upward tendency.

Table 2
Descriptive statistics of lexical complexity

Metrics	Group	*Mean*	*SD*	*n*
vocd-D at Time 1	EG	39.038	29.939	60
	CG	32.476	25.416	68
vocd-D at Time 2	EG	42.122	23.584	60
	CG	30.616	22.480	68
vocd-D at Time 3	EG	40.567	15.266	60
	CG	39.089	14.483	68
vocd-D at Time 4	EG	45.136	15.778	60
	CG	43.752	14.020	68

The Levene's test was not significant for vocd-D ($p \geq 0.358$). The test of between-subjects effects revealed that the main effect of the group factor on the average score of vocd-D across time was statistically significant ($F_{(1, 125)} = 3.945$, $p = 0.049 < 0.05$) with a small-to-medium effect size ($\eta^2_p = 0.031 > 0.01$). Partial eta-squared (η^2_p) was employed to estimate the effect size, and the thresholds suggested are small $\eta^2_p > 0.01$; medium $\eta^2_p > 0.06$; large $\eta^2_p > 0.14$ (Cohen, 1988; Miles & Shevlin, 2001). Therefore, the EG students performed significantly better over time than their CG counterparts in terms of lexical complexity.

Syntactic complexity

Table 3 shows the descriptive statistics of the three metrics of syntactic complexity in each group at each time point. The estimated marginal means of overall complexity and phrasal complexity showed a roughly V-shaped and increasing pattern. The estimated marginal means of subordination complexity demonstrated an evidently upward tendency.

Table 3
Descriptive statistics of syntactic complexity

Metrics	Group	*Mean*	*SD*	*n*
OvComp at Time 1	EG	8.086	2.288	68
	CG	7.390	1.924	60
OvComp at Time 2	EG	4.573	1.652	68
	CG	3.863	0.958	60
OvComp at Time 3	EG	10.031	3.167	68
	CG	9.423	2.934	60
OvComp at Time 4	EG	14.061	5.490	68
	CG	11.164	3.504	60
PhrComp at Time 1	EG	5.811	1.237	68
	CG	5.374	0.878	60
PhrComp at Time 2	EG	3.523	1.049	68
	CG	2.993	0.554	60
PhrComp at Time 3	EG	5.958	1.298	68
	CG	5.759	1.387	60
PhrComp at Time 4	EG	6.550	1.712	68
	CG	6.054	1.531	60
SubComp at Time 1	EG	1.387	0.253	68
	CG	1.362	0.203	60
SubComp at Time 2	EG	1.287	0.160	68
	CG	1.286	0.181	60
SubComp at Time 3	EG	1.677	0.390	68
	CG	1.625	0.329	60
SubComp at Time 4	EG	2.187	0.863	68
	CG	1.878	0.553	60

Note. OvComp = overall complexity, PhrComp = phrasal complexity, SubComp = subordination complexity

The Levene's test on overall complexity was significant on Time 2 ($p = 0.034 < 0.05$) and Time 4 ($p = 0.013 < 0.05$). For phrasal complexity, the Levene's test was significant on Time 2 ($p = 0.002 < 0.05$). For subordination complexity, the result of Levene's test on Time 4 was significant ($p = 0.039 < 0.05$). Nevertheless, a violation of the equal variance assumption is less of an issue with roughly equivalent sample sizes because the ratio of the larger sample size ($n = 68$) to the smaller one ($n = 60$) is less than the threshold of 1.5 (Pituch & Stevens, 2016). The test of between-subjects effects showed that the main effect of the group factor on the average score of overall complexity across time was statistically significant ($F_{(1, 125)} = 12.338$, $p = 0.001 < 0.05$) and the effect size was medium to large ($\eta^2_p = 0.09 > 0.06$). On the other hand, for the specific metrics of syntactic complexity, a significant difference of the main effect of the group factor was detected on the average score of phrasal complexity ($F_{(1, 125)} = 8.108$, $p = 0.005 < 0.05$) and subordination complexity ($F_{(1, 125)} = 5.147$, $p = 0.025 < 0.05$) across time, respectively, and their effect size was of medium magnitude ($\eta^2_p = 0.061$ for phrasal complexity and $\eta^2_p = 0.04$ for subordination complexity). In a nutshell, from the perspective of overall complexity, the EG students produced significantly more words per AS-unit than their counterparts in the CG. In terms of phrasal complexity and subordination complexity respectively, the EG students generated significantly more words per clause and significantly more clauses per AS-unit than their CG counterparts.

Within-subjects effects of time on students' oral language complexity

Lexical complexity

The results of Mauchly's test of sphericity showed that sphericity was not assumed ($p < 0.001$) for the vocd-D value and the corresponding Greenhouse-Geisser epsilon was 0.795, higher than the threshold of 0.75 (O'Brien & Kaiser, 1985). Therefore, the Huynh-Feldt adjustment with the univariate tests was used. The tests of within-subjects effects showed that the main effect of time was not statistically significant on the average scores on the vocd-D value ($F_{(2.475, 309.351)} = 0.483$, $p = 0.658 > 0.05$), sphericity not assumed. Additionally, the group × time interaction effects on vocd-D were also not statistically significant ($F_{(2.475, 309.351)} = 2.720$, $p = 0.055 > 0.05$), sphericity not assumed. In other words, the time factor did not lead to

any statistically significant effects on lexical complexity and there was also no statistically significant group × time interaction effect on lexical complexity.

Syntactic complexity

The results of Mauchly's test of sphericity showed that sphericity was not assumed for any of the three metrics of syntactic complexity ($p < 0.001$). The Greenhouse-Geisser epsilon for overall complexity was 0.670 and that for subordination complexity was 0.542, both less than 0.75. For phrasal complexity, the Greenhouse-Geisser epsilon was 0.875, greater than 0.75. Therefore, the Greenhouse-Geisser adjustment with the univariate tests was used for overall complexity and subordination complexity, while the Huynh-Feldt adjustment was used for phrasal complexity. Accordingly, the tests of within-subjects effects showed that the main effect of the time factor was not statistically significant on the average scores of overall complexity ($F_{(2.010, 251.3)} = 0.577$, $p = 0.563 > 0.05$) or subordination complexity ($F_{(1.627, 203.314)} = 0.587$, $p = 0.523 > 0.05$), sphericity not assumed. However, the main effect of the time factor was indeed statistically significant on the average scores of phrasal complexity ($F_{(2.73, 341.229)} = 3.152$, $p = 0.029 < 0.05$), sphericity not assumed (Figure 5) and the effect size was small to medium ($\eta^2_p = 0.025 > 0.01$). In other words, the students in each group seemed to produce significantly more words per clause (phrasal complexity) on average over time. However, the average number of words per AS-unit (overall complexity) or the average number of clauses per AS-unit (subordination complexity) did not seem to change significantly over time.

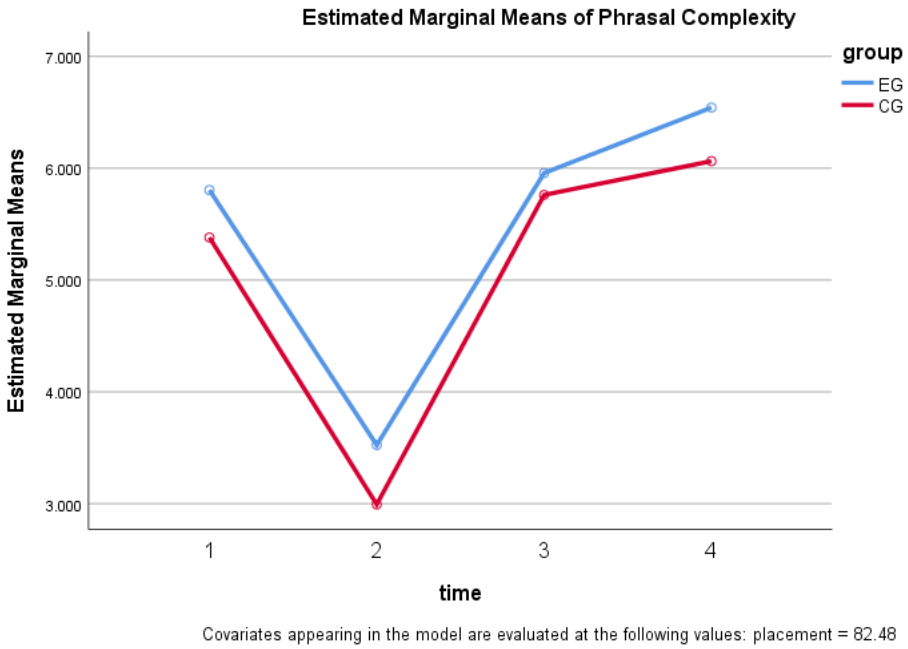


Figure 5. Profile plots of phrasal complexity

Moreover, the group × time interaction effect on phrasal complexity was not statistically significant ($F_{(2.73, 341.229)} = 0.555$, $p = 0.629 > 0.05$), whereas the group × time interaction effects on overall complexity ($F_{(2.010, 251.3)} = 5.507$, $p = 0.001 < 0.05$) and on subordination complexity ($F_{(1.627, 203.314)} = 4.160$, $p = 0.023 < 0.05$) were indeed statistically significant (Figures 6 and 7). The effect size was small-to-medium ($\eta^2_p = 0.042$ for overall complexity and $\eta^2_p = 0.032$ for subordination complexity, both greater than 0.01). In a nutshell, the time factor had a statistically significant effect on phrasal complexity but no group × time interaction effect was found on phrasal complexity. Conversely, the time factor did not lead to any statistically significant effects on overall complexity or subordination complexity, but significant group × time interaction effects were found on the two metrics.

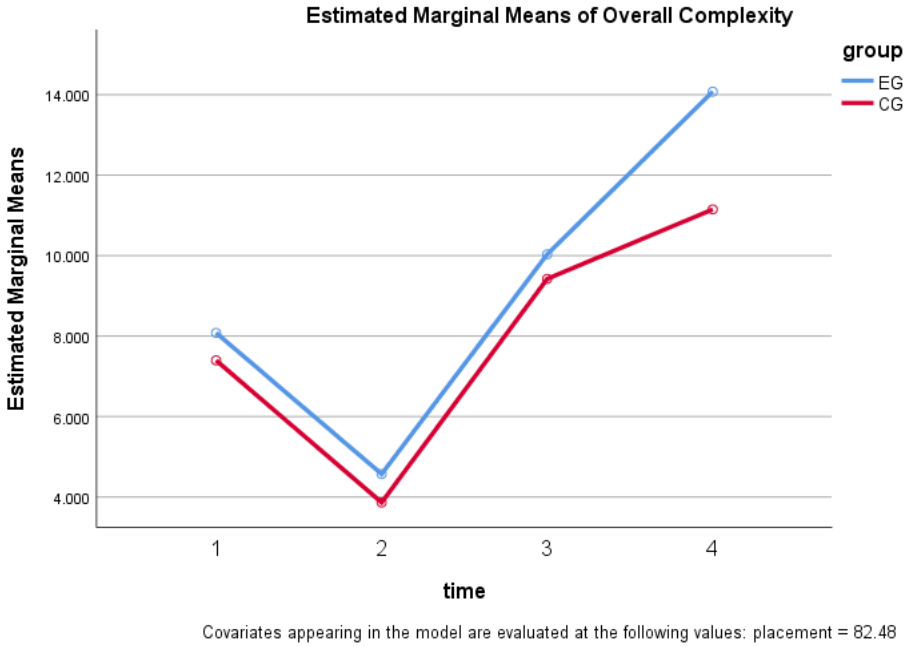


Figure 6. Profile plots of overall complexity

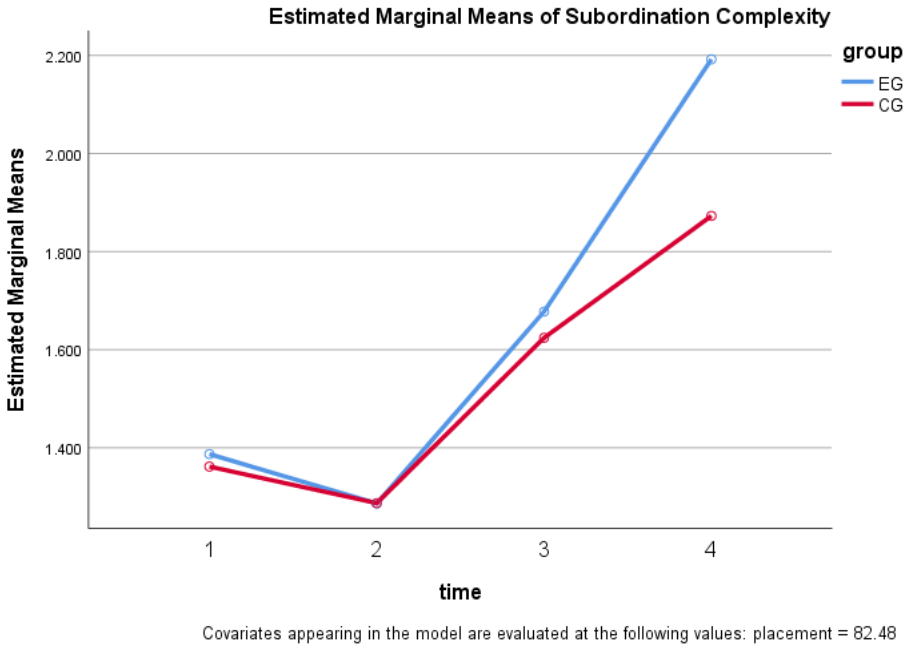


Figure 7. Profile plots of subordination complexity

Accordingly, follow-up simple-effects tests were conducted to describe the nature of the significant group × time interaction effects on overall complexity and subordination complexity. Results of the multivariate tests by group (Table 4) showed that for the EG students, there existed a statistically significant difference in the mean differences in overall complexity scores (Wilks' lambda = 0.189, $p < 0.001$) and subordination complexity scores over time (Wilks' lambda = 0.436, $p < 0.001$). Likewise, for the CG students, there also existed a statistically significant difference in the mean differences in overall complexity scores (Wilks' lambda = 0.246, $p < 0.001$) and subordination complexity scores over time (Wilks' lambda = 0.624, $p < 0.001$).

Table 4
Multivariate tests on overall complexity and subordination complexity by group

Metrics	Group	Wilks' lambda	*F*	Hypothesis *df*	Error *df*	Sig.	Partial η^2	Noncent. parameter	Observed power[b]
OvComp	EG	.189	175.614[a]	3.000	123.000	.000	.811	526.842	1.000
	CG	.246	125.719[a]	3.000	123.000	.000	.754	377.157	1.000
SubComp	EG	.436	53.109[a]	3.000	123.000	.000	.564	159.326	1.000
	CG	.624	24.724[a]	3.000	123.000	.000	.376	74.172	1.000

Note. Each *F* tests the multivariate simple effects of time within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.
a. Exact statistic
b. Computed using alpha = .05

Pairwise tests of mean differences in overall complexity between time points within each group were conducted with Bonferroni as the adjustment for multiple comparisons. Results showed that for the EG students, there existed a significant difference between each of the time points ($p < 0.001$) (Table 5 & Figure 6). For the CG students, there also existed a significant difference between each time point ($p \leq 0.043$) (Table 5 & Figure 6).

Table 5
Pairwise comparisons on overall complexity between time points by group
Measure: overall complexity

Group	(I) time	(J) time	Mean difference (I–J)	SE	Sig.[a]	95% confidence interval for difference[a]	
						Lower bound	Upper bound
EG	1	2	3.509*	.271	.000	2.782	4.236
		3	-1.951*	.375	.000	-2.956	-.947
		4	-5.994*	.571	.000	-7.524	-4.464
	2	3	-5.461*	.355	.000	-6.412	-4.509
		4	-9.503*	.534	.000	-10.935	-8.070
	3	4	-4.042*	.593	.000	-5.631	-2.453
CG	1	2	3.532*	.289	.000	2.758	4.305
		3	-2.026*	.399	.000	-3.095	-.957
		4	-3.753*	.608	.000	-5.382	-2.123
	2	3	-5.557*	.378	.000	-6.570	-4.544
		4	-7.284*	.569	.000	-8.809	-5.759
	3	4	-1.727*	.631	.043	-3.418	-.035

Note. Based on estimated marginal means
a. Adjustment for multiple comparisons: Bonferroni.
* The mean difference is significant at the .05 level.

Then, pairwise tests of mean differences in overall complexity between groups at each time point were also conducted with Bonferroni as the adjustment. Results showed that for Time 2 and Time 4, there existed statistically significant differences between the two groups (for Time 2, $p = 0.004 < 0.05$; for Time 4, $p = 0.001 < 0.05$, Table 6). Figure 6 depicts the differentiated trending of overall complexity by the two cohorts of students. The EG students had a greater gradient than their CG counterparts at the ending phase of the course (i.e., from Time 3 to Time 4).

Table 6
Pairwise comparisons on overall complexity between groups by time point

Measure: overall complexity							
						95% confidence interval for difference[b]	
Time	(I) group	(J) group	Mean difference (I–J)	SE	Sig.[a]	Lower bound	Upper bound
Time 1	EG	CG	.685	.378	.072	-.063	1.433
Time 2	EG	CG	.707*	.244	.004	.224	1.190
Time 3	EG	CG	.610	.545	.265	-.468	1.689
Time 4	EG	CG	2.926*	.830	.001	1.284	4.568

Note. Based on estimated marginal means
a. Adjustment for multiple comparisons: Bonferroni.
* The mean difference is significant at the .05 level.

Likewise, for subordination complexity, pairwise tests of mean differences between time points within each group were conducted with Bonferroni as the adjustment for multiple comparisons. Results showed that there existed a statistically significant difference between each of the four time points for the EG students. By contrast, for the CG students, there existed a statistically significant difference between some time points but not all (e.g., Time 1–Time 3, Time 1–Time 4; see Table 7).

Table 7
Pairwise comparisons on subordination complexity between time points by group

Measure: subordination complexity							
						95% confidence interval for difference[a]	
Group	(I) time	(J) time	Mean difference (I-J)	SE	Sig.[a]	Lower bound	Upper bound
EG	1	2	.101*	.028	.003	.025	.176
		3	-.290*	.048	.000	-.419	-.162
		4	-.805*	.092	.000	-1.051	-.560
	2	3	-.391*	.048	.000	-.521	-.261
		4	-.906*	.086	.000	-1.135	-.676
	3	4	-.515*	.093	.000	-.763	-.267
CG	1	2	.075	.030	.087	-.006	.155
		3	-.263*	.051	.000	-.399	-.126
		4	-.511*	.098	.000	-.773	-.250
	2	3	-.075	.030	.000	-.155	.006
		4	-.337*	.052	.000	-.475	-.199
	3	4	.337*	.052	.078	.199	.475

Note. Based on estimated marginal means
a. Adjustment for multiple comparisons: Bonferroni.
* The mean difference is significant at the .05 level.

Then, pairwise tests of mean differences in subordination complexity between groups at each time point were conducted with Bonferroni as the adjustment. Results showed that for Time 4, there existed statistically significant differences between the two groups ($p = 0.015 < 0.05$, Table 8). Figure 7 illustrates the differentiated trending of subordination complexity by the two cohorts. Similar to the pattern of overall complexity, the EG students had a greater gradient than their CG counterparts at the ending phase of the course (i.e., from Time 3 to Time 4).

Table 8
Pairwise comparisons on subordination complexity between groups by time point

Measure: subordination complexity							
						95% confidence interval for difference[a]	
Time	(I) group	(J) group	Mean difference (I-J)	SE	Sig.[a]	Lower bound	Upper bound
Time 1	EG	CG	.025	.041	.537	-.056	.107
Time 2	EG	CG	-.001	.030	.982	-.060	.059
Time 3	EG	CG	.053	.065	.410	-.074	.181
Time 4	EG	CG	.319*	.130	.015	.063	.576

Note. Based on estimated marginal means
a. Adjustment for multiple comparisons: Bonferroni.
* The mean difference is significant at the .05 level.

To recap, the findings revealed above were recapitulated below (Table 9), summarising the main effects of the group factor and the time factor, as well as the group × time interaction effects. Moreover, the trending of each measure was illustrated with headed arrows.

Table 9
Summary of the findings

Measures	RQ1: between-subjects effects	RQ2: within-subjects effects	RQ3: interaction effects	trending
lexical complexity	(sm)	×	×	↑
phrasal complexity	(m)	(sm)	×	$\uparrow^{v}$
overall complexity	(ml)	×	(sm)	$\uparrow^{v}$
subordination complexity	(sm)	×	(sm)	$\uparrow^{v}$

Note. × denotes a non-significant effect; (m) = medium effect size, (sm) = small-to-medium effect size, (ml) = medium-to-large effect size; ↑ = an upward tendency, $\uparrow^{v}$ = a V-shaped upward tendency.

Discussion

Effects of harnessing ASR technology on oral complexity

The EG students' significant gains on complexity revealed by the findings from RQ1 indicate that the use of the ASR technology in pre-class self-learning had positive effects on English oral complexity. The ASR-based mediating tasks before class provided language learners with an avenue for interacting with the application with immediate feedback. The success of flipped classrooms is heavily dependent on students' preparedness and pre-class self-learning (M. Y. C. Jiang et al., 2020). When self-learning the flipped content before class, the students often come across some content knowledge beyond their understanding. Therefore, immediate feedback from reliable sources is a crucial element for an effective self-learning in a flipped setting. To some extent, the ASR technology can play such role in English oral practice in a flipped EFL classroom. With the aid of the ASR technology, students' linguistic performance, such as their oral complexity, can be significantly enhanced, which puts the tool beyond merely a means of drill and practice.

Specifically, the ASR technology seemed to assist the EG students in managing their discourse flow in a more effortless and attention-free fashion (Tavakoli et al., 2016). As mentioned, learners' repetitious practice on the pre-class mediating tasks may lead to a degree of proceduralisation in speaking in the target language. Such proceduralisation may further lead language learners to develop a state of automatisation in oral expression and might help them speak more effortlessly in the target language (DeKeyser, 2001, 2007; Segalowitz, 2010). Consequently, the students could have more attentional resources allotted to the complexity of their oral output due to the integration of the ASR technology. The significant between-subjects difference unveiled in RQ1 evidenced the positive effects of the ASR technology in a flipped EFL classroom. However, since the ASR technology is yet to be widely used in language classrooms, more empirical studies are needed to examine the connections between the utilisation of the ASR technology and the proceduralisation more systematically.

Enhanced oral complexity over time

The findings from RQ2 revealed that regardless of group membership, students' phrasal complexity was significantly enhanced over a semester from Time 2 to Time 4. This finding is consistent with those of some previous CAF-related empirical studies (e.g., Lennon, 1990; Serrano., 2012), indicating that even though they did not perform mediating tasks in pre-class self-learning, the CG students still improved significantly on phrasal complexity as their counterparts in the EG did. Compared with overall complexity and subordination complexity, which are both based on AS-unit, phrasal complexity is only a clause-based metric. Since an AS-unit consists of at least one clause (Foster et al., 2000), the structure of a clause may be less complicated than an AS-unit. Therefore, it might be easier for the students to improve their syntactic complexity at the phrasal level rather than at the clausal or sentential level.

Despite the significantly enhanced phrasal complexity across the two groups, the developmental trajectories were roughly parallel between each time point, with no significant interaction effect. The lack of variation in gradient might be due to the clause length constraints in spoken language (Chafe, 1988; Vercellotti, 2019) because the students must balance between increasing syntactic complexity and making sure the meaning could get through to the other interlocuters (Vercellotti, 2019). Figure 5 depicts a roughly parallel set of trajectories of the two cohorts, in a V-shaped upward pattern with Time 2 as the minimum. Such pattern echoes the other two metrics of syntactic complexity (Figures 6 and 7), indicating that the in-class task at Time 2 (i.e., Unit 4) might be restrictive on the development of students' oral proficiency at the syntactic complexity level due to certain systematic task elements. Further reflections and investigations are needed to clarify the reasons from a design-based perspective.

Moreover, follow-up simple-effects tests showed that overall complexity and subordination complexity were enhanced significantly in the two respective groups (Table 4). Therefore, improvements on all the three metrics of syntactic complexity were witnessed over time. In contrast, students' lexical complexity was not enhanced significantly over time. These contrasting findings align with most CAF studies conducted among non-native speakers (e.g., Kalantari & Gholami, 2017; Skehan, 2009a). Those studies found a negative correlation between lexical complexity and syntactic complexity, indicating that for the non-native speakers, "more varied lexis seems to cause problems for non-native speakers and provokes more errors while not driving forward complexity" (Skehan, 2009a, p. 116).

Interaction effects of group × time on overall complexity and subordination complexity

The findings from RQ3 revealed a similar trending at the end of the programme (i.e., from Time 3 to Time 4) on both overall complexity and subordination complexity: before Time 3, the developmental trajectories of overall complexity and subordination complexity were in a parallel pattern and there was little significant between-group difference, except for the overall complexity at Time 2 ($p = 0.004 < 0.05$). From Time 3 to Time 4, however, the gradient of the EG trajectory seemed to be greater than that of the CG (Figures 6 and 7). Moreover, at Time 4, the EG students outperformed their counterparts in the CG on overall complexity ($p = 0.001 < 0.01$) and subordination complexity ($p = 0.015 < 0.01$). This indicated that at the ending phase of the program (i.e., Time 3 through Time 4), a faster development was witnessed in the EG on overall complexity and subordination complexity. In contrast, there was no such interaction effect on phrasal complexity, indicating that integrating the ASR technology resulted in slower but significantly more gains in syntactic complexity at the sentential level other than at the phrasal level. This finding corroborates some previous studies (e.g., De Clercq & Housen, 2017).

Inconsistent with those involving L2 written data and uncovering a levelling off or decrease in subordination towards the upper proficiency levels (see the synthesis in Norris & Ortega, 2009), the present study, based on spoken data, found that the overall and subordination complexity of the EG developed even faster. The variation in the gradients of the developmental trajectories after Time 3 seemed to be closely associated with the integration of ASR technology. Evidently, it took some time for this developmental difference to be manifested significantly. As mentioned, through the practice of the mediating tasks, the EG students could achieve some degree of proceduralisation. According to Ortega (2003), during the process of proceduralisation, beginning and intermediate FL or L2 learners may prefer complexity by subordination. In contrast, phrasal complexity may be favoured at more advanced L2 proficiency. Therefore, some researchers in the field of L2 writing (e.g., Biber et al., 2016; Housen et al., 2019; Kyle & Crossley, 2018) have proposed using more fine-grained measures that address different types of syntactic

complexity to gauge the potential changes in complexification at the sentential level. Besides, since the gradient change occurred only at the Time 3–Time 4 interval, more empirical evidence is needed.

Conclusion and limitations

With respect to the between-subjects effects, the EG students performed statistically better on lexical complexity and all the metrics of syntactic complexity than their counterparts in the CG. To be specific, the EG students articulated more complex utterances at the lexical level (lexical complexity) than their counterparts in the CG. Likewise, the EG students also produced more words per AS-unit (overall complexity), more words per clause (phrasal complexity) and more clauses per AS-unit (subordination complexity) than their counterparts in the CG.

Concerning the within-subjects effects, significant improvement in phrasal complexity was witnessed over time in both groups. In contrast, overall complexity and subordination complexity did not seem to change significantly over time. Conversely, in light of the group × time interaction effects, significant interaction effects were witnessed on overall complexity and subordination complexity. The gradients of the EG trajectories of the two metrics from Time 3 to Time 4 were greater than those of the CG. However, on phrasal complexity, the interaction effect was not significant.

There are several limitations in this study that need to be addressed in future research. First, the participants were only enrolled in one public university in China, which could result in a lower representativeness of the sample. Therefore, more studies in different flipped tertiary EFL settings are needed to explore the effects of integrating ASR technology on students' oral English complexity. Moreover, follow-up studies should be conducted from a perspective of design-based research (Jong et al., 2021) to explore further and confirm our findings. Also, more empirical studies need to be conducted to evaluate students' peer interaction in addition to oral complexity, such as accuracy, fluency and task-based interactional features.

Second, qualitative data need to be collected to interpret and triangulate our findings. Post-intervention interviews and classroom observations could be employed to conduct mixed-method research and provide a holistic picture of the developmental trajectories of students' English oral complexity.

Third, due to the regulation of the university, we were not allowed to collect data in person in the classroom. Therefore, in this study, the students had to record their own performance of the in-class tasks. It became harder to ensure the quality and authenticity of the recordings, which might result in a data loss.

Acknowledgements

Special thanks to the Chinese University of Hong Kong Stanley Ho Big Data Decision Analytics Research Centre and the Chinese University of Hong Kong Teaching Development and Learning Enhancement Grant awarded to Professor Helen Meng for supporting this interdisciplinary research.

References

Akçayır, G., & Akçayır, M. (2018). The flipped classroom: A review of its advantages and challenges. *Computers & Education, 126*, 334–345. https://doi.org/10.1016/j.compedu.2018.07.021

Albert, A. (2011). When individual differences come into play: The effect of learner creativity on simple and complex task performance. In P. Robinson (Ed.), *Second language task complexity* (pp. 239–265). John Benjamins Publishing Company. https://doi.org/10.1075/tblt.2.16ch9

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman.

Bachman, L. (2005). *Statistical analysis for language assessment.* Oxford University Press.

Bax, S., Nakatsuhara, F., & Waller, D. (2019). Researching L2 writers' use of metadiscourse markers at intermediate and advanced levels. *System, 83*, 79–95. https://doi.org/10.1016/j.system.2019.02.010

Bergmann, J., & Sams, A. (2012). *Flip your classroom: Reach every student in every class every day.* International Society for Technology in Education.

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, *37*(5), 639–668. https://doi.org/10.1093/applin/amu059

Blair, E., Maharaj, C., & Primus, S. (2016). Performance and perception in the flipped classroom. *Education and Information Technologies*, *21*, 1465–1482. https://doi.org/10.1007/s10639-015-9393-5

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.32.02bul

Bulté, B., Housen, A., Pierrard, M., & Van Daele, S. (2008). Investigating lexical proficiency development over time: The case of Dutch-speaking learners of French in Brussels. *Journal of French Language Studies*, *3*(18), 277–298. https://doi.org/10.1017/S0959269508003451

Chafe, W. (1988). Punctuation and the prosody of written language. *Written Communication*, *5*(4), 395–426. https://doi.org/10.1177/0741088388005004001

Chapelle, C. (2003). *English language learning and technology: Lectures on applied linguistics in the age of information and communication technology*. John Benjamins Publishing Company. https://doi.org/10.1075/lllt.7

Chien, S. Y., Hwang, G. J., & Jong, M. S. Y. (2020). Effects of peer assessment within the context of spherical video-based virtual reality on EFL students' English-speaking performance and learning perceptions. *Computers & Education*, *146*, Article 103751. https://doi.org/10.1016/j.compedu.2019.103751

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

DeClercq, B., & Housen, A. (2017). A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. *The Modern Language Journal*, *101*(2), 315–334. https://doi.org/10.1111/modl.12396

DeKeyser, R. (2001). Automaticity and automatization. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 125–151). Cambridge University Press. https://doi.org/10.1017/cbo9781139524780.007

DeKeyser, R. (2007). Situating the concept of practice. In R. DeKeyser (Ed.), *Practicing in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 1–18). Cambridge University Press. https://doi.org/10.1017/cbo9780511667275.002

Ellis, N., & Robinson, P. (2008). An introduction to cognitive linguistics, second language acquisition, and language instruction. In P. Robinson & N. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 3–24). Routledge. https://doi.org/10.4324/9780203938560

Ellis, R. (2003). *Task-based language learning and teaching.* Oxford University Press.

Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford University Press.

Evers, K., & Chen, S. (2020). Effects of an automatic speech recognition system with peer feedback on pronunciation instruction for adults. *Computer Assisted Language Learning.* https://doi.org/10.1080/09588221.2020.1839504

Foster, P., Tonkyn, A., Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons, *Applied Linguistics*, *21*(3), 354–375. https://doi.org/10.1093/applin/21.3.354

Fox, W. H., & Docherty, P. D. (2019). Student perspectives of independent and collaborative learning in a flipped foundational engineering course. *Australasian Journal of Educational Technology*, *35*(5), 79–94. https://doi.org/10.14742/ajet.3804

Franco, H., Bratt, H., Rossier, R., Rao Gadde, V., Shriberg, E., Abrash, V., & Precoda, K. (2010). EduSpeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, *27*(3), 401–418. https://doi.org/10.1177/0265532210364408

Geng, J., Chai, C. S., Jong, M. S. Y., & Luk, E. T. H. (2019). Understanding the pedagogical potential of interactive spherical video-based virtual reality from the teachers' perspective through the ACE framework. *Interactive Learning Environments.* https://doi.org/10.1080/10494820.2019.1593200

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2012). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, *27*(1), 70–105. https://doi.org/10.1080/09588221.2012.700315

Horwitz, E. (2002). Language anxiety and achievement. *Annual Review of Applied Linguistics*, *21*, 112–126. https://doi.org/10.1017/S0267190501000071

Housen, A., De Clercq, B., Kuiken, F., & Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second language Research*, *35*(1), 3–21. https://doi.org/10.1177/0267658318809765

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 1–20). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.32.01hou

Jiang, M. Y. C., Jong, M. S. Y., Lau, W. W. F., Chai, C. S., Liu, K. S. X., & Park, M. (2020). A scoping review on flipped classroom approach in language education: challenges, implications and an interaction model. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2020.1789171

Jiang, Y., & Dewaele, J. M. (2020). The predictive power of sociobiographical and language variables on foreign language anxiety of Chinese university students. *System, 89*, Article 102207. https://doi.org/10.1016/j.system.2020.102207

Jong, M. S. Y. (2015). Context-aware geography field trip with EagleEye: Teachers' first experience. In M. Chang, & Y. Li (Eds.), *Smart learning environments* (pp. 77–93). Springer. https://doi.org/10.1007/978-3-662-44447-4_5

Jong, M. S. Y. (2017). Empowering students in the process of social inquiry learning through flipping the classroom. *Educational Technology & Society, 20*(1), 306–322. https://drive.google.com/file/d/14I9vx3wTvxQIeNNvwZGEX8JbJFZBdfL0/view

Jong, M. S. Y. (2019). To flip or not to flip: Social science faculty members' concerns about flipping the classroom. *Journal of Computing in Higher Education, 31*(2), 391–407. https://doi.org/10.1007/s12528-019-09217-y

Jong, M. S. Y., Chan, T., Tam, V., & Jiang, M. Y. C. (2021). Design-based research on gamified outdoor social enquiry learning with context-aware technology: Integration of teacher facilitation for advancing the pedagogical effectiveness. *International Journal of Mobile Learning & Organisation, 15*(1), 107–126. https://doi.org/10.1504/IJMLO.2021.111601

Jong, M. S. Y., Chen, G. W., Tam, V., & Chai, C. S. (2019). Adoption of flipped learning in social humanities education: The FIBER experience in secondary schools. *Interactive Learning Environments, 27*(8), 1222–1238. https://doi.org/10.1080/10494820.2018.1561473

Jong, M. S. Y., Lee, J. H. M., & Shang, J. J. (2013). Educational use of computer game: Where we are and what's next? In R. Huang, Kinshuk, & J. M. Spector (Eds.), *Reshaping learning: Frontiers of learning technology in a global context* (pp. 299–320). Springer. https://doi.org/10.1007/978-3-642-32301-0_13

Jong, M. S. Y., Shang, J. J., Lee, F. L., Lee, J. H. M., & Law, H. Y. (2006). Learning online: A comparative study of a game-based situated learning approach and a traditional web-based learning approach. In Z. Pan, R. Aylett, H. Diener, X. Jin, S. Gobel, & L. Li (Eds.), *Lecture notes in computer science: Vol. 3942. Technologies for e-learning and digital entertainment* (pp. 541–551). Springer. https://doi.org/10.1007/11736639_65

Kalantari, R., & Gholami, J. (2017). Lexical complexity development from dynamic systems theory perspective: Lexical density, diversity, and sophistication. *International Journal of Instruction, 10*(4), 1–18. https://doi.org/10.12973/iji.2017.1041a

Kyle, K., & Crossley, S. E. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal, 102*(2), 333–349. https://doi.org/10.1111/modl.12468

Lan, Y. J., Botha, A., Shang, J. J., & Jong, M. S. Y. (2018). Technology-enhanced contextual game-based language learning. *Educational Technology & Society, 21*(3), 86–89. https://drive.google.com/file/d/1k-qkKwklE5pKfkREsf0CJhm2nKQWoHoM/view

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, Instruments, & Computers, 41*(3), 841–849. https://doi.org/10.3758/BRM.41.3.841

Lee, G., & Wallace, A. (2018). Flipped learning in the English as a foreign language classroom: Outcomes and perceptions. *TESOL Quarterly, 52*(1), 62–84. https://doi.org/10.1002/tesq.372

Lennon, P. (1990). The advanced learner at large in the L2 community: Developments in spoken performance. *International Review of Applied Linguistics in Language Teaching, 28*, 309–324. https://doi.org/10.1515/iral.1990.28.4.309

Levis, J. (2007). Computer technology in teaching and researching pronunciation. *Annual Review of Applied Linguistics, 27*, 184–202. https://doi.org/10.1017/S0267190508070098

Lin, H. C., & Hwang, G. J. (2019). Research trends of flipped classroom studies for medical courses: A review of journal publications from 2008 to 2017 based on the technology-enhanced learning model. *Interactive Learning Environments*, *27*(8), 1011–1027. https://doi.org/10.1080/10494820.2018.1467462

Lin, H. C. S., Yu, S. J., Sun, J. C. Y., & Jong, M. S. Y. (2019). Engaging university students in a library guide through wearable spherical video-based virtual reality: Effects on situational interest and cognitive load. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2019.1624579

McCarthy, P. M., & Jarvis, S. (2007) vocd: A theoretical and empirical evaluation. *Language Testing, 24*, 459–488. https://doi.org/10.1177/0265532207080767

McCarthy, P. M, & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McCrocklin, S. M. (2016). Pronunciation learner autonomy: The potential of automatic speech recognition. *System*, *57*, 25–42. https://doi.org/10.1016/j.system.2015.12.013

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, *15*(3), 323–337. https://doi.org/10.1093/llc/15.3.323

Miles, J., & Shevlin, M. (2001). *Applying regression and correlation: A guide for students and researchers*. Sage.

Moranski, K., & Henery, A. (2017). Helping learners to orient to the inverted or flipped language classroom: Mediation via informational video. *Foreign Language Annals*, *50*(2), 285–305. https://doi.org/10.1111/flan.12262

Nakazawa, K. (2012). The effectiveness of focused attention on pronunciation and intonation training in tertiary Japanese language education on learners' confidence: Preliminary report on training workshops and a supplementary computer program. *The International Journal of Learning*, *18*(4), 181–192. https://doi.org/10.18848/1447-9494/cgp/v18i04/47590

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555–578. https://doi.org/10.1093/applin/amp044

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, *97*, 316–333. https://doi.org/10.1037/0033-2909.97.2.316

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college level L2 writing. *Applied Linguistics*, *24*(4), 492–518. https://doi.org/10.1093/applin/24.4.492

Ozudogru, M., & Aksu, M. (2020). Pre-service teachers' achievement and perceptions of the classroom environment in flipped learning and traditional instruction classes. *Australasian Journal of Educational Technology*, *36*(4), 27–43. https://doi.org/10.14742/ajet.5115

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601. https://doi.org/10.1093/applin/amp045

Pituch, K. A., & Stevens, J. P. (2016). *Applied multivariate statistics for the social sciences* (6th ed.). Routledge.

Rahman, A. A., Aris, B., Rosli, M. S., Mohamed, H., Abdullah, Z., & Zaid, N. M. (2015). Significance of preparedness in flipped classroom. *Advanced Science Letters*, *21*(10), 3388–3390. https://doi.org/10.1166/asl.2015.6514

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732942

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics*, *45*, 237–257. https://doi.org/10.1515/iral.2007.009

Segalowitz, N. (2010). *Cognitive basis of second language fluency*. Routledge. https://doi.org/10.4324/9780203851357

Serrano, R. (2012). A longitudinal analysis of the effects of one year abroad. *Canadian Modern Language Review*, *68*(2),138–163. https://doi.org/10.3138/cmlr.68.2.138

Skehan, P. (1996). Second language acquisition research and task-based instruction. In J. Willis & D. Willis (Eds.), *Challenge and change in language teaching* (pp. 17–30). Heinemann.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Skehan, P. (2003). Task based instruction. *Language Teaching*, *36*(1), 1–14. https://doi.org/10.1017/S026144480200188X

Skehan, P. (2009a). Lexical performance by native and non-native speakers on language learning tasks. In B. Richards, H. M. Daller, D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition: The interface between theory and application* (pp. 107–124). Palgrave Macmillan. https://doi.org/10.1057/9780230242258_7

Skehan, P. (2009b). Modelling second language performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics*, *30*(4), 510–532. https://doi.org/10.1093/applin/amp047

Song, Y., Jong, M. S. Y., Chang, M., & Chen, W. (Eds.). (2017). "HOW" to design, implement and evaluate the flipped classroom? ‑A synthesis [Editorial]. *Educational Technology & Society, 20*(1), 180–183. https://drive.google.com/file/d/1N9l7yxSdwKk7pbP-Pnguog-ubWj2o1BN/view

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, *60*(2), 1–46. https://doi.org/10.1111/j.1467-9922.2010.00562.x

Steel, C. H., & Levy, M. (2013). Language students and their technologies: Charting the evolution 2006-2011. *ReCALL*, *25*(3), 306–320. https://doi.org/10.1017/S0958344013000128

Stockwell, G. (2008). Investigating learner preparedness for and usage patterns of mobile learning. *ReCALL*, *20*(3), 253–270. https://doi.org/10.1017/S0958344008000232

Sun, Z. R., & Xie, K. (2020). How do students prepare in the pre-class setting of a flipped undergraduate math course? A latent profile analysis of learning behavior and the impact of achievement goals. *The Internet and Higher Education*, *46*, 1–13. https://doi.org/10.1016/j.iheduc.2020.100731

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 133–150. https://doi.org/10.1515/iral-2016-9994

Tavakoli, P., Campbell, C., & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic intervention. *TESOL Quarterly*, *50*(2), 447–471. https://doi.org/10.1002/tesq.244

Thorndike, E. (1932). *The fundamentals of learning.* AMS Press.

Tseng, J. J., Chai, C. S., Tan, L., & Park, M. (2020). A critical review of research on technological pedagogical and content knowledge (TPACK) in language teaching. *Computer Assisted Language Learning.* https://doi.org/10.1080/09588221.2020.1868531

Vercellotti, M. L. (2019). Finding variation: Assessing the development of syntactic complexity in ESL Speech. *International Journal of Applied Linguistics*, *29*(2), 233–247. https://doi.org/10.1111/ijal.12225

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*(1), 65–83. https://doi.org/10.1177/026553220001700103

Corresponding author: Morris Siu-Yung Jong, mjong@cuhk.edu.hk

Please cite as: Jiang, M. Y.-C., Jong, M. S.-Y., Lau, W. W.-F., Chai, C.-S., & Wu, N. (2021). Using automatic speech recognition technology to enhance EFL learners' oral language complexity in a flipped classroom. *Australasian Journal of Educational Technology*, *37*(2), 110–131. https://doi.org/10.14742/ajet.6798

Appendix: A sample in-class task

1. Pre-class mediating task (Unit 4): *Talking about Memories*
Work by yourself and talk to iFlyRec (讯飞听见) on your smart device. Talk as much as you can about:

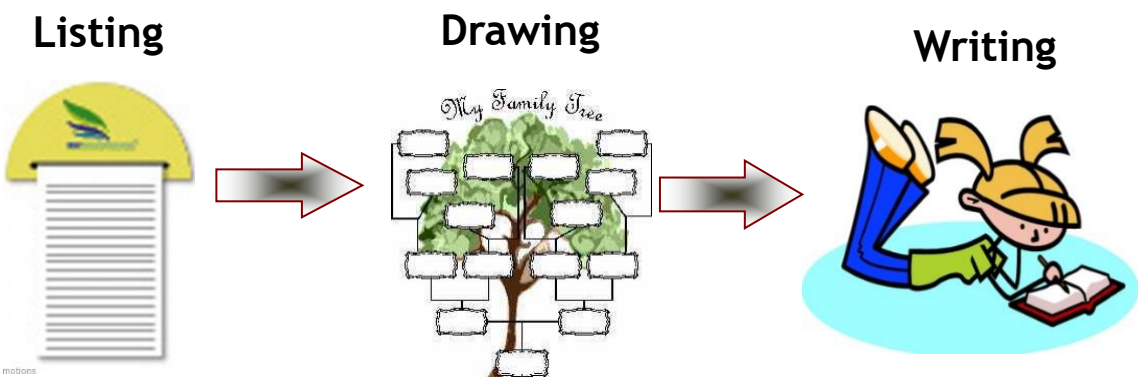
- a person or an object which you associate with your family, or
- an occasion when you realized how much your parents loved you, or
- how you felt when you started college

Try to use the feedback provided by the app and repeat your practice. Once you can keep speaking fluently on this task for at least one minute, upload your audio file as an attachment via Unipus.

2. In-class communicative task (Unit 4): *Talking about Your Family*

Make a list of the people in your family, and draw lines between them to show the relationship. Talk about each person and make sure you include some details about:

- their name
- their age
- their nickname (if they have one)
- their character
- how well you get on with them
- any favourite or typical stories about them



Work in pairs and exchange roles when you are ready.

Student A: Talk about three of your family members (excluding parents) to Student B.
Student B: Try and decide which is Student A's closest and favourite family member.