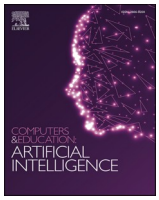
Computers and Education: Artificial Intelligence 3 (2022) 100044

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence



Verifying student identity in oral assessments with deep speaker

Jake Renzella [a,*], Andrew Cain [b], Jean-Guy Schneider [b]

[a] School of Computer Science and Engineering, University of New South Wales, Sydney, NSW, 2052, Australia
[b] School of Information Technology, Deakin University, Geelong, Victoria, Australia

ARTICLE INFO

Keywords:
Academic integrity
Artificial intelligence
Machine learning
Audio feedback
Learning management systems

ABSTRACT

Contract cheating, a form of academic misconduct in which students outsource assessment activities to third parties, is a topic of concern among educators. As similarity-detection systems are ineffective at detecting contract cheating, some institutions have turned to intensely criticised proctoring systems, however student and educator bodies report high costs and privacy concerns. Oral assessment is an alternative assessment approach that provides valuable interpersonal and communication skills in graduates and can naturally help detect and deter cheating. However, oral assessment is typically time-consuming, and in larger courses, it is challenging to validate respondents' identity. Advancements in machine learning approaches can scale time-consuming tasks that previously required prohibitive educator effort. One such system, Deep Speaker, is a speaker identification and verification system that can verify if two audio samples resemble speech from the same person with high accuracy. This paper presents an innovative tool that integrates an online oral assessment tool, Real Talk, with Deep Speaker. This proposed system facilitates scalable student-tutor discussions while providing longitudinal student identity validation with minimal cost and impact for institutions and addressing student privacy concerns. We evaluated the system and showed that student audio responses collected via oral discussion tools are suitable for verification. We then discuss the impact our system may have when applied in higher education. We posit that institutions can use such approaches to detect cases of contract cheating, enhance learning outcomes, and pave the way for more student-friendly assessment and discussion models in online education.

1. Introduction

Teaching and learning software systems may utilise audio tools for various reasons, most prominently to support the facilitation of online oral assessment, or the provision of formative feedback to students, for which the literature indicates a variety of benefits when used in tandem with written feedback (Parkes & Fletcher, 2017; Gleaves & Walker, 2013; Wood et al., 2011; Merry & Orsmond, 2008). Recent advances in the web browser ecosystem, such as the deployment of the WebRTC API in major browsers (GoogleInc, 2019) further fosters the development of user-friendly audio features to support student learning; no longer requiring the use of third-party browser extensions or file managers. These increasingly accessible platforms encouraged our research team to develop an audio feedback system into the open-source Doubtfire learning and feedback tool. The Doubtfire learning and feedback tool is a research-driven system designed to facilitate a teaching and learning approach based on the tenets of constructive alignment, and cyclical formative feedback (Renzella & Cain, 2020). The discussion and audio-discussions tools available in the Doubtfire tool enable educators to provide formative feedback alongside assessments, and for students to respond, iterate their work, and re-submit for further assessment.

Growing online cohorts (Palvia et al., 2018), increasing concerns of academic integrity (Velliaris, 2016; Lancaster & Clarke, 2016, pp. 639–654), as well as the impact of COVID-19 on higher education (Vlachopoulos, 2020), has motivated educators to seek alternative forms of assessment for higher education. One growing area of concern of educators is contract cheating behaviour. Contract Cheating, a term first coined by Clark and Lancaster (Clarke & Lancaster, 2006), is used to describe a form of academic misconduct or cheating wherein students engage third-parties such as friends, family, or professional services to produce original pieces of work. Because of the originality of the work produced, plagiarism detection systems are not effective. Investigations in reliably detecting cases of contract cheating indicate that educators may be able to detect cases by leveraging their experience, knowledge and relationship with their students (Harper et al., 2019). Dawson and Sutherland-Smith (2018) present that educators can detect contract

* Corresponding author.
E-mail addresses: jake.renzella@unsw.edu.au (J. Renzella), andrew.cain@deakin.edu.au (A. Cain), jeanguy.schneider@deakin.edu.au (J.-G. Schneider).

https://doi.org/10.1016/j.caeai.2021.100044
Received 29 July 2021; Received in revised form 12 November 2021; Accepted 12 December 2021
Available online 20 December 2021

cheating at the marking stage with sufficient training. Other techniques which are becoming the focus of discussion include the use of discussion or examinations of students *viva voce*, to ascertain if contract cheating activities had potentially taken place (Lancaster & Clarke, 2016, pp. 639–654). In response, the audio tools in Doubtfire were expanded upon and developed into the Real Talk tool, which supports viva voce-style assessment and discussions (Renzella, Cain, & Schneider, 2021) in an approach that scales with larger cohorts.

Real Talk is a student-synchronous, teacher-asynchronous discussion tool used to present students with audio prompts in order to discuss specific aspects of completed assessment tasks. Real Talk ensures that the student responds to the tutor's prompts in real-time, providing an authentic response capturing the student's understanding of their work. This work aims to provide a demonstration and initial evaluation of an end-to-end approach to both solicit authentic student responses with Real Talk, and to apply a speaker verification system, such as Deep Speaker, to safeguard against the capability of students to bypass this approach by soliciting others to respond to their Real Talk discussion prompts. With promising initial results (Renzella, Tubino, et al., 2021), we hope that institutions can use similar student-friendly, intelligent teaching assistant systems (Yacef, 2002) to pave the way for alternative student engagement and assessment models, illuminate opportunities for meaningful learning interventions, all while addressing concerns of contract cheating and general academic misconduct.

This paper is organised as follows: Section 1 presents the introduction, Section 2 includes the background and related work. Section 3 presents the research questions and desired outcomes. Section 4 describes the system components and the approach under evaluation. Section 5 presents the research design, followed by the results in Section 6. Section 7 discusses the results, and the impact the approach has on students, teachers and pedagogy. Section 8 concludes the paper.

2. Background and related work

Academic misconduct such as plagiarism and contract cheating have been long-standing concerns of educators. The rising popularity of online education (Department of Education and Training, 2018; Barnes & Paris, 2013), the impact of COVID-19 (García-Peñalvo et al., 2021; Kumar, 2020), and the increasing accessibility and use of contract cheating services (Evans, 2018) have motivated universities to reevaluate and approach online assessment in new and innovative ways. Motivations and opportunities which drive student cheating activities are varied. A large-scale Australia study conducted by Bretag et al. (2019) reported dissatisfaction with teaching quality, opportunity, the stakes of the assessment and many more as factors which seem to correlate with student cheating activities.

Naturally, academic institutions utilise a variation of a written examinations to combat academic misconduct in higher education; however, recent studies call into question the security of the written exam. Harper et al. (2019) surveyed students of eight Australian universities and found that students reported that cheating activities were more prevalent in exams than assignments, most commonly in multiple-choice, followed by short-answer questions. The large-scale study's results are contrary to the widespread belief that the exam format is more secure than other forms of assessment. The lowest reported exam type in which students reported receiving assistance was the oral or viva exam.

Reinforcing these concerns is the reported prevalence of contract cheating behaviours. Media reports and investigations yield varying results, some indicating that up to 15% of students admit to buying essays (Evans, 2018). While the actual rates of contract cheating may be impossible to discover, the consensus is that the behaviour is increasing as the prevalence and persuasion of these services increase (Newton, 2018; Rowland et al., 2018).

To combat the rising concerns surrounding contract cheating, institutions have turned to technology. Online examination proctoring is a prevalent approach in Australia for ensuring academic integrity in assessment activities. Online examination proctoring involves the use of software systems to monitor a variety of markers or behaviours, including the microphone, webcam, IP address, browser activity or network activity to detect suspicious behaviour (O'reilly & Creagh, 2016). Some proctoring systems such as Examinity[1] support live human invigilation throughout the online assessment, while others such as Proctorio[2] allow an invigilator to examine an integrity report following the completion of the assessment. Students have strongly protested these proctored systems, citing the systems as being "intrusive" and raising privacy concerns (Haskell-Dowland, 2020; Australian Broadcasting Corporation, 2020). Further, mandatory activities such as ensuring that students sweep their workspace with a video camera before commencing an assessment activity may introduce a stereotype threat (Spencer et al., 2016) which could elucidate problematic behaviours such as reduced performance, and requires active student participation which disrupts the assessment activity taking place.

One promising response to the academic integrity concerns is the use of oral examinations of students (*viva voce*). Educators can use oral assessment to discuss submitted assignments or engage in general evaluations of understanding. Further, the discussions themselves can help ascertain if contract cheating had potentially taken place (Lancaster & Clarke, 2016, pp. 639–654; Renzella, Cain, & Schneider, 2021), as well as providing general summative assessment. Akimov and Malin (2020) presents a case study evaluating the application of oral assessment as an online assessment tool, and utilising the Joughin (1998) *matrix of oral examination validity, reliability and fairness*, concludes that while there are some limitations of online oral assessment, the approach is beneficial, and teaches valuable interpersonal and communication skills in graduates. While at the same time, "practically eliminating cheating in the assessment".

Students also perceive oral assessment as a more secure method of assessment. Harper et al. (2019) surveyed Australian higher education students and found that students identified oral defence as one of four categories of assessment that are the least likely to be outsourced (contract cheated) by fellow students. Later, Harper et al. (2021) reported that oral assessment saw the lowest rate of student-reported exam assistance. Lancaster and Clarke (2016, pp. 639–654) also support these claims, while presenting two significant disadvantages of *viva voce* assessment which may inhibit more widespread adoption: the time-consuming nature of individual discussions, including scheduling activities, and fairness inconsistencies of the difficulty of questions provided to a cohort. These disadvantage, primarily the former, is commonly echoed by others (Akimov & Malin, 2020; Walker & Townley, 2012).

Real Talk is an audio-based discussion tool designed to provide the benefits of real-time, student-tutor discussions, while addressing these disadvantages. The tool collects a series of discussion prompts from a tutor which ask students to discuss aspects of their completed assignments. Once collected, the audio prompts are stored, and only made available to the student when they begin their discussion with the system. The design of the system ensures students can not access the prompt (s) before they are engage in the audio-based discussion. The user interface of Real Talk is shown in Fig. 1, and presents the two main student views. Left, shows the microphone and speaker testing phase, which ensures the student verifies that their hardware is functioning correctly before gaining access to the discussion prompts. The right view shows the student listening to a prompt and responding in real time. The audio visualisation in the center represents the student's audio throughout the discussion, and the countdown on the right presents the remaining time available for response. Once the final prompt is responded to, the student can finish the discussion and the resulting

[1] https://www.examity.com/.
[2] https://proctorio.com/.

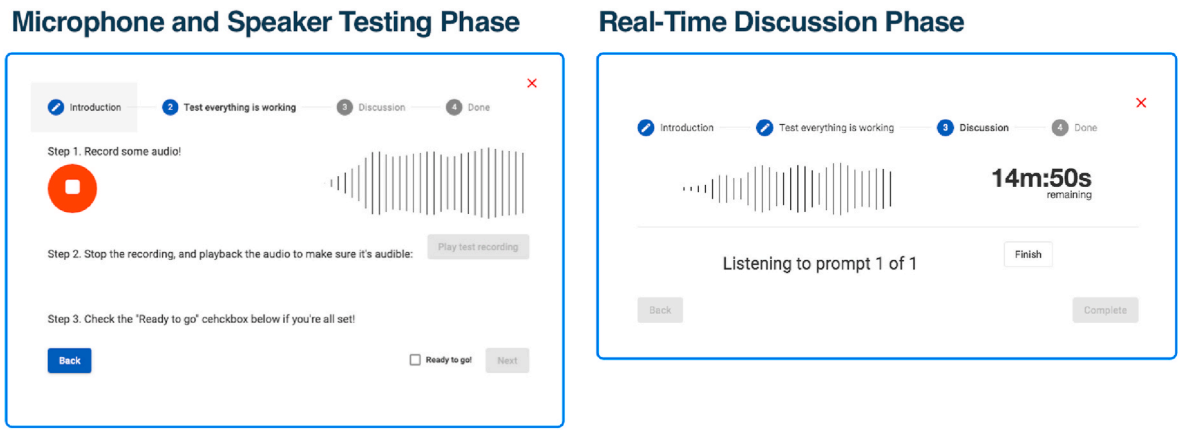


Fig. 1. The Real Talk system user interface, showing the mandatory hardware testing stage (left), and the main discussion stage (right).

audio including the tutor prompts, and student response in a single audio file is sent to the server.

The Real Talk process is presented in more detail in Section 2. An Australian university deployed the tool and approach to approximately 40 computing courses during COVID-19-affected teaching periods and found that the real-time nature of the student's engagement with the tool ensured educators could trust that the student's response represented their understanding, and resulted in meaningful learning interventions which supported student learning outcomes (Renzella, Cain, & Schneider, 2021). While Real Talk was positively received as a discussion tool, there were concerns raised regarding the lack of student identity verification.

Due to Real Talk's tutor-asynchronous engagement model, the tool supports educators in conducting online oral assessment without requiring scheduling activities. Due to this asynchronous nature however, there is less opportunity for educators to verify the identity of students (for example, in an online discussion via a video call software, the tutors could ask the student to present their ID card which isn't possible in Real Talk). In this paper we explore whether a speaker verification system such as Deep Speaker can be applied to audio collected via the Real Talk system, address previous shortcomings of oral examinations, and introduce verification of student identity over time. This is all towards designing and evaluating a tool/model for scalable, secure, online oral discussion/examination.

2.1. Intelligent agents in teaching and learning

When considering approaches for designing and integrating artificial intelligence-based systems in education, there are a number of approaches available. A description of these are presented in Table 1. For the sake of this paper, we take the philosophy encompassed in the approach of Intelligent Teaching Assistant (ITA) systems. Intelligent Teaching Assistants (Yacef, 2002) support the tutor in their ongoing activities. The reason we take this approach with non-deterministic artificial intelligence systems such as Deep Speaker, is that when these systems perform poorly, the tutor is left to make the final decision. This tutor-in-the-loop approach ensures that students are not negatively impacted by system-related issues.

Table 1
Approaches to intelligent agents in teaching and learning systems.

Approach/System	Tailored to	Description
Pedagogical Agents	Learner	Often animated, anthropomorphic, deliver content proactively
Intelligent Tutoring System	Learner	Deliver content proactively, authoritative
Intelligent Teaching Assistant	Learner & teacher	Facilitates the teaching and learning system, non-authoritative
Learning Companion System	Learner	Non-authoritative

2.2. Deep speaker

Deep Speaker is a neural speaker embedding system designed to facilitate speaker recognition tasks (Li et al., 2017). Deep Speaker, first published in 2017, marked a dramatic improvement in the accessibility of state-of-the-art speaker recognition tasks, lowering the Equal Error Rate by 30% and error rate (false positive or false negative rate) by 50% compared to previous state-of-the-art approaches. Deep Speaker supports a range of speaker recognition tasks include speaker verification and speaker identification, for both text dependent and independent spoken phrases. Speaker verification uses a speakers voice to verify a claim that a speaker is a specific identity. Speaker identification algorithms attempt to identify an unknown speaker's identity from a list of identities. Both speaker verification and identification activities may be useful in the context of maintaining academic integrity in higher education. For example, this paper explores the use of speaker verification to verify the identity of a student when engaging in oral assessment or discussions. Speaker identification could be used to identify secondary speakers in an audio sample (for example, to ensure educator's speech does not verify a student's identity). Text dependent activities require the speaker to repeat a specific phrase for enrollment and verification activities, whereas text-independent activities do not have this requirement. The benefits of text-independent verification is that the process does not require additional student-activities (such as reading a verification phrase before beginning assessment), however is less accurate than text-dependent activities).

3. Research questions

Broadly, this research project seeks to investigate the feasibility of applying a speaker verification system (in this case the Deep Speaker system's speaker verification algorithms) on student audio responses captured via a learning and feedback tool. The study sets out to determine if biases and design approaches in the training dataset (described in Section 4.1) would make the model unsuitable for use in verifying student identity from audio discussion or assessment responses. The research questions and outcomes are listed below:

Research Question 1. *Can machine-learning speaker verification systems (such as Deep Speaker) verify the identity of students throughout their engagement in oral discussion and assessment tools?*
Research Question 2. *How could speaker verification systems and student-tutor oral discussion and assessment tools such as Real Talk be integrated?*

Research Question 3. *What are the implications and limitations of applying discussion and verification systems such as Real Talk and Deep Speaker in higher education settings?*

The motivating goal behind RQ1 is to evaluate the feasibility of introducing the Deep Speaker verification system into the Real Talk system. Fig. 2 presents the Business Process Modelling Notation diagram of the proposed additions to the Real Talk system, with the inclusion of the custom Deep Speaker-based system in the context of the Doubtfire learning and feedback tool shown in the BPMN swimlane titled "Deep Speaker". Previously, discussion responses were sent directly to the system for the tutor to review. After the addition of Deep Speaker to the Real Talk system, once the student audio is captured during the discussion process the audio is sent to the Python-based program which wraps Deep Speaker. The system handles the anonymised enrolment of student audio into an interim profile. If the audio is the first example for the user, a new profile is generated and the associated audio is enrolled for that profile. Subsequent audio for the same user is then verified against the enrolled audio with the confidence score returned for the educator to review.

4. System and data

The audio dataset used for evaluation in this study was collected via the audio tools in the Doubtfire learning and feedback tool. The dataset was isolated to only include asynchronous audio comments which helped to ensure a sufficiently large dataset, and ensure only student audio responses were included in the analysis. This limitation meant that audio responses from Real Talk discussions (which contain both tutor prompt(s) alongside student responses) were not included in the analysis.

This project utilised Deep Speaker's text-independent speaker verification capability. The process is divided into two stages; enrollment and verification on a pre-trained model.

4.1. Model training data

The Deep Speaker machine-learning model used in this study was trained on the openly available LibriSpeech Automatic Speech Recognition Corpus (Panayotov et al., 2015), a large-scale (1,000 h) corpus of English speech with 2,484 individual speakers. The body of speech was collected by recording speakers reading text.

4.2. Speaker enrollment

The enrollment phase takes an identifier of a student profile, and stores a de-identified ID for that student in an SQLite database. For this research project, the original user ID associated with the audio recordings was also stored so that manual analysis of the original audio examples were possible. In production systems, the identifiable user id is not required to be stored, which ensures the end-to-end verification process is entirely non-identifiable. This is especially important for data collection and storage policies of sensitive student data. The associated audio is then enrolled by generating an associated Mel-frequency Cepstrum (MFCC) and storing the MFCC data in the database alongside the de-identified user ID. As stated, subsequent examples for the same user were then generated into an MFCC, and verified against the stored profile.

4.3. Speaker verification

Fig. 2 describes the verification process in the bottom-most lane of the diagram. Any time student audio is collected via audio discussion or feedback tools, the audio example is sent to the Deep Speaker system alongside the student ID. The ID is then used to determine if there has been a previous enrollment associated with that student. If this is the case, the Mel-frequency Cepstrum (MFCC) of that enrollment is retrieved from the database, and an MFCC for the audio being analysed is generated. The Deep Speaker model is used to determine if the two MFCCs belong to the same speaker. The result is then sent back to the Real Talk System displayed to the user. Due to the reduced feature-set of

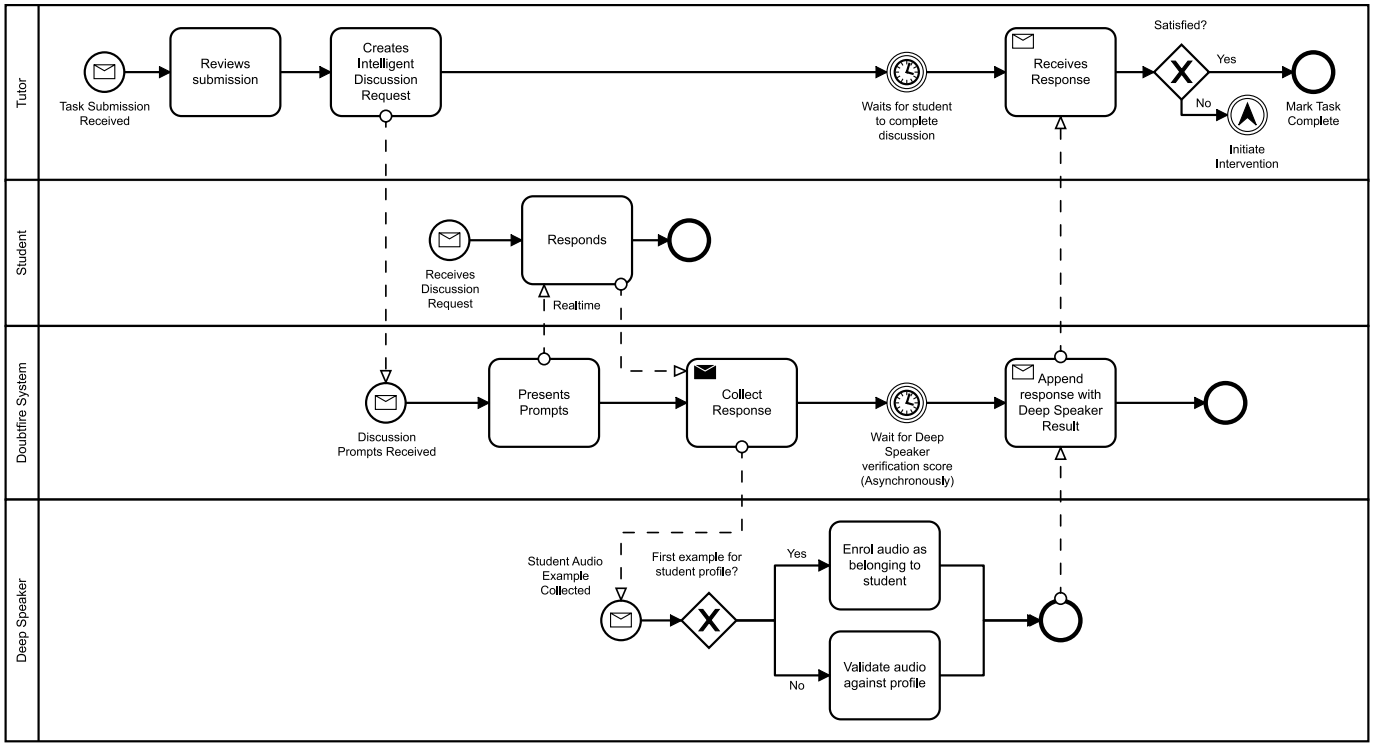


Fig. 2. BPMN diagram of the Real Talk system, describing the audio collection process, and Deep Speaker verification process.

the MFCC compared to the standard audio examples, the verification process is quick, returning a result to be viewed by the tutor.

4.4. Pre-processing

As enrollment and verification algorithms only use small amounts of audio, our custom pipeline ensures all audio submitted for enrollment or verification is pre-processed to determine the segment of the audio most likely to contain clear speech data. This process determines the segments of the audio with the highest energy levels, and clips the audio to that segment. If no segment is able to be identified, audio can not be used in analysis. In this context, energy levels refers to a measurement of the intensity of audio. Fig. 3 presents a visualisation of how the segment of audio is determined, as shown, the highlighted region is processed as it is the first segment of audio which exceeds the energy threshold.

The speaker verification pipeline requires audio to be converted and processed in mono format in the pre-processing phase, with a sample rate of 16 kHz (Renzella & Griffiths, 2021). Therefore, the audio recording pipeline should record at a sample rate of at least 16 kHz, however recordings can be made in a variety of formats or channels and can simply be converted to mono 16 KHz during inference.

5. Methodology

To validate the Deep Speaker model performance on student speech, we evaluated the model against audio comments submitted by students. The dataset was collected over a period of four years via the Doubtfire learning and feedback tool in a variety of Information Technology units in an Australian university. The corpus composed primarily of student audio responses to discussion prompts, and formative feedback provided by tutors. The inclusion criteria consisted of ensuring:

- **a) audio must be from a student account** - this is to ensure the study was focused in its evaluation and to ignore biases which may be present in non-student speech,
- **b) more than one example of audio must exist for the student** - this is to ensure there is sufficient speech from two separate events to use for enrollment and analysis,
- **c) the audio examples must include sufficient audio data and be of a minimum duration (minimum 3 s)** - this is to ensure sufficient data for the enrollment and verification algorithms.

This quantitative study seeks to evaluate the accuracy of deep speaker by calculating the sensitivity (True Positive Rate TPR defined as the TP/TP + FN), and overall accuracy to evaluate the classification model, similar to (Li et al., 2017). determining two key metrics: a) the performance of the model in successfully verifying that two samples of speech were from the same student, and b) performance of the model in successfully identifying that two samples of speech originated from different students. For clarity, we declare positive and negative tests with the following definitions:

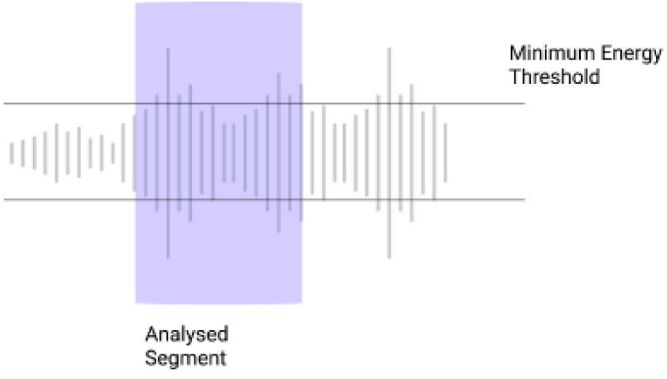


Fig. 3. A visualisation of how the segment of audio to be analysed is determined.

- **true positive** - The model correctly identified that two samples came from the same speaker
- **false positive** - The model incorrectly reported that the two samples did come from the same speaker
- **true negative** - The model correctly identified that two samples did not come from the same speaker
- **false negative** - The model incorrectly reported that two samples were not from the same speaker

For each series of audio examples belonging to a student, the first example of audio was enrolled into the Deep Speaker system with the resulting MFCC data stored in a local database. Subsequent audio examples known to come from the same student profile were then validated, this is shown as check 1 in the Fig. 4. In order to validate the *true positive* rate a random selection of the high-confidence evaluations (confidence score greater than 50) were manually reviewed to verify the results were not *false positives*. Following this, a subset of all verification scores lower than 50% were then marked for manual review to determine if the evaluation was a *false negative* or if there were issues with the audio itself. The manual analysis involved listening to all examples of audio associated with that user and determining if the listener agreed with the model's outcome, or believed the examples were from the same speaker. If the manual review disagreed with the model's result, we identified what aspects of the audio may have caused the disagreement. To validate the *true negative* rate, we cross-checked a sample of audio from different student profiles to ensure the model returned a score which indicated that the identity of the speakers was not consistent (less than 50) shown as check 2 of Fig. 4. A sample of the responses were manually reviewed to ensure we agreed with the model response, including all that returned a positive value (testing for *false positives*).

By calculating the true-positive rate, we can calculate the overall accuracy of the performance of the model, and validate the approach's suitability for verifying identity in student audio responses.

6. Results

The dataset accessed consisted of 10,274 examples of student audio from 33 units ranging a variety of undergraduate and postgraduate courses. After removing examples which did not meet the inclusion criteria, the resulting dataset consisted of 2,448 audio samples associated with 658 student profiles. On average, each student profile consisted of 3.7 audio samples, giving one for enrolment and an average of 2.7 samples for verification.

Deep Speaker enrollment and verification was then applied to each series, resulting in 1,790 verifications taking place for check 1. Of these, 1,450 (82%) of verifications returned a score greater than 50%, with 340 reporting low confidence that the audio was from the same speaker. A subset of these verifications were manually reviewed to determine the occurrences of false negatives. Of the samples with a result which indicated that the identity was consistent, a subset was randomly selected to manually review for false positives. No cases of false positives were identified.

The most common cause of low-scoring results were due to issues with the underlying audio. In almost all of these cases the reviewer agreed with the model's output, or concluded that the audio was insufficient for verifying the consistency of the identity of the speaker. As the dataset was not curated, there were instances where the audio enrolled or verified consisted only of static noise, or non-speech data. In some cases, the student profile was enrolled with this low-quality audio, causing all subsequent verifications to return a low-scoring result. In these cases we marked these as true-negatives.

In 5% of cases, the reviewer disagreed with the model's verification

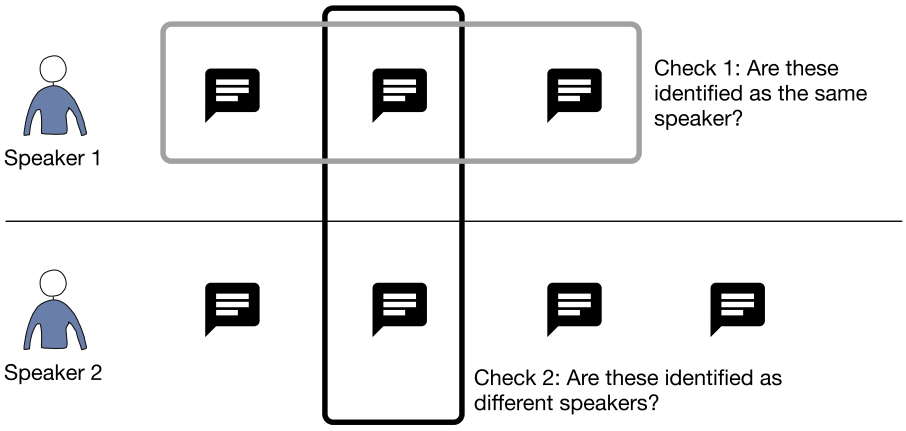


Fig. 4. The two checks performed to evaluate suitability of Deep Speaker.

score. There was a single case where the identity of the speaker appears to have changed between two recordings, this was detected by the audio verification system, which the review team agreed with and was therefore marked as a true-negative.

For check 2, we conducted 1,050 verifications across different student profiles and found 24 occurrences where the model indicated that the speaker's identity was the same (2.4% false positive rate). These 24 instances were also manually reviewed and determined to be false positives.

We calculated the true-positive rate, as well as total accuracy of the Deep Speaker model and compared it to the accuracy generated by the original training audio dataset (Rémy, 2020). The accuracies are presented in Table 2.

7. Discussion

RQ.1 sought to determine if student audio collected in the Doubtfire learning and feedback tool in an Australian University performed similarly to the Deep Speaker model trained on a corpus of speech generated by collecting participant recordings of spoken text. The results indicate that the system performed in line with the model's described performance when the audio was of sufficient quality and isolated to the student speaker. Further, the system performed well at distinguishing between the identity of different speakers, indicating an acceptable false-positive rate. While Deep Speaker's overall high performance in verifying students' identity may have been expected, discussing the implications of using such a system in practice is essential in ensuring such approaches can be successfully integrated into higher education. The remaining sections of this paper explore the impact that the introduction of verified, oral assessment may have on students, educators and pedagogy.

7.1. Capturing student audio for verification tasks

Any artificial intelligence system that intends to use audio for identity verification (such as Real Talk with Deep Speaker) requires speech of sufficient clarity, isolation and duration. If the enrollment audio is of insufficient quality, then all future verification instances will return low confidence results. This requirement highlights the importance of having high-quality, trustworthy enrollment data. In this study, we assumed that the audio sample used for enrolment contained the student's voice, which was not always the case. Several approaches can strengthen the audio collection and verification process before allowing students to submit audio in discussion tools.

Speech-to-Text tools may offer one way for audio tools to ensure enrolment data include speech. Rather than only verifying that collected audio is not silent, enhancements such as running a Speech-to-Text algorithm on collected audio can verify that speech is of sufficient quality to be detected by machine learning models. While this would not verify the identify of the speaker, it could be used to allow the software to automatically enrol user with a greater confidence that the audio captured will be suitable for verification activities.

Higher quality enrolment data, with greater authority in relation to identity of the speaker, could be obtained if audio enrolment was incorporated within institutional enrolment processes. If institutions want to rely on such systems to broadly address academic integrity concerns, a human-invigilated audio sample could be collected and enrolled when the student joins an institution. For example, when students present sufficient identification and are given their student identification card, an audio sample could be captured and enrolled. All future interactions could use that sample as the source of truth, thereby providing both a high quality audio enrolment but also one that is known to have come from the student themselves.

There are legal implications and requirements which may need consideration before such a system can be implemented. While these issues are regional and change depending on the country or city, considerations of ownership of likeness in audio recordings, two-party consent regulations, and privacy laws may need to be considered. In this study, ethics approval was obtained for analysis of audio recordings which were knowingly submitted by students.

Table 2
A breakdown of the Deep Speaker's performance when applied to student speech.

Analysis Type	TPR[a]	Accuracy
Deep Speaker Sample Evaluation	73%	99.6%
Student-Isolated Speech	92%	95%

[a] True-Positive Rate (Sensitivity).

7.2. Multiple voices

The analysis presented in this paper utilised isolated student audio containing responses to formative feedback provided by a tutor in the Doubtfire system. When applying the Deep Speaker system to Real Talk responses (which may contain both tutor prompts and student responses in a single audio file) a few considerations will need to be taken to ensure that the correct speaker is enrolled and verified. Where one audio sample contains more than a single voice, the segments of the audio used for verification activities will determine which speaker is processed.

The Deep Speaker algorithm operates on small segments of audio, so it possible different strategies may need to be applied to identify the different speakers within a sample. Our implementation looked only at the first segment of audio likely to contain speech, and therefore only identified the first speaker. This impacted analysis of Deep Speaker on

end-to-end Real Talk discussion responses (which contained both tutor prompts and student responses in a single piece of audio). The results indicated a performance accuracy of around 18%. This low performance is due to the likelihood that the tutor's audio may be detected and verified against the student's enrollment profile or otherwise interfering with speaker recognition tasks. This limitation also applies to the analysis of recorded discussion calls over an online audio/video calling platform.

To address this problem, speaker verification systems such as that described in this paper must attempt to isolate portions of the audio to identify segments expected only to include student speech. Real Talk is provided information on when the student's response is being appended to the discussion recording. In the future, Real Talk will generate time-markings on the audio that can be used to isolate student speech for enrollment or analysis. Speaker identification (determining the number of speakers within a sample) could also be determined, though the value of this will depend on whether or not the audio will be listened to by a human. In our Real Talk use case, the audio is expected to be listened to by the student's tutor and therefore this was seen as low value and was not explored further.

7.3. Ensuring trust in the AI system

An important aspect of this system is the human-in-the-loop approach to artificial intelligence-based interventions. Due to the non-deterministic nature of machine-learning systems such as Deep Speaker, educators need to ensure that the possibly erroneous outcomes of inferences do not negatively impact students, tutors, or reduce the trust in AI systems. In our case where a student's verification score is calculated to be low (indicating the system does not believe the speaker's identity is consistent), the system's only course of action is to notify the student's tutor for manual review and intervention. Designing the system in this way (as an Intelligent Teaching/Tutoring Assistant) is important for ensuring trust in the system from the perspective of both students and teaching staff.

Determining how to identify the point at which the confidence result is deemed unacceptably low is unclear. For this study, the 50% threshold was chosen, however we are investigating further techniques to both a) better identify appropriate performance thresholds as the underlying speaker recognition models may change (Cummaudo et al., 2020), and b) to decrease the likelihood of low-quality audio being submitted (such as validating the audio with speech to text-to-speech algorithms to ensure the speech is clear).

7.4. Academic integrity impact

While plagiarism is detectable with similarity systems for most examples of student work, contract cheating is much more challenging to identify. Rather than trying to detect cases of plagiarism or contract cheating outright, the approach encapsulated by the Real Talk system aims to evaluate the student's understanding of their submitted work through discussions. Several design choices have been made to reduce the inclination and capability of students to cheat the Real Talk assessment itself. The Real Talk tool aims to address academic integrity concerns through:

- **System Design:** The tool limits student access to tutor's audio prompts so as to assess the student's authentic and immediate response. The Real Talk system is designed to imitate an in-person audio discussion, without requiring scheduling one-to-one interview times but with limitations on the ability to clarify or provide follow up questions. Getting an immediate response ensures an authentic representation of the student's current understanding of their work.
- **Oral Assessment:** Real Talk is an oral assessment tool. Students report cheating the least in oral assessments compared to any other assessment methods (Harper et al., 2021).
- **Small, low-stakes assessments:** As discussed in section 1, students are reportedly less likely to engage in academic misconduct in low-stakes activities (activities which are not associated with significant portions of their final grade). Real Talk discussions were initially designed to accompany a series of small, low-stakes assessment tasks (Renzella & Cain, 2017). Real Talk discussions did not directly contribute to student grades, but supported the associated submission and could be used to ask students to redo work where concerns in understanding were identified.
- **Multiple discussions across a unit:** The low-overheads required for a tutor to create a Real Talk discussion means that the tool can be used with some regularity within a unit. In our practice, we typically see two to three Real Talk discussions per student per unit (Renzella, Tubino, et al., 2021). If a student were to attempt to contract cheat these discussions, they would be required to engage multiple times.
- **Identity verified by Deep Speaker:** Following the promising analysis of Deep Speaker in this study, the difficulty for students to contract out their Real Talk discussions grows. While a student could contract their Real Talk discussions to a third party, they would need to ensure that the same speaker responds in discussions from all units throughout their degree. If the enrollment process did become a secure activity overseen by the institution, then this would introduce further challenges for cheating the system.
- **Understanding verified by a tutor:** Finally, an important aspect of Real is that it is designed for tutors to listen to student responses and identify both potential gaps in understanding and learning interventions. Suppose the student engaged in academic misconduct with the underlying task or assignment (but not the Real Talk discussion); in that case, the tutor may detect this if the student does not understand the work they have submitted. Even if the student is not engaging in any academic misconduct, they will still benefit from these discussions with their tutor. In our practice, we train our tutors to provide prompts about work already submitted, so the discussions encompass more than just knowledge recall. (For example, "How did you come up with this solution? and what would happen if you completed it with solution X?).

7.5. Observations following use of Real Talk

Over approximately the past two years, we have utilised the Real Talk tool in the Doubtfire learning and feedback tool in an Australian university, without the Deep Speaker verification component. The positive outcomes we have experienced after using the system are described in (Renzella, Tubino, et al., 2021), however we have also identified a number of limitations and challenges which are relevant when incorporating the Deep Speaker component. Firstly, the system does not aim to detect cases of plagiarism where the student can still provide an understanding of their submitted work. Therefore similarity systems are still required. Live coaching is also an area of concern, wherein students sitting a Real Talk discussion may have assistance coaching them (while muting their microphone hardware) on how to respond to the prompts. Currently, tutors may detect this by identifying periods of silence in the response following the prompt. Future work aims to automatically identify suspicious periods of silence in the Real Talk responses. Pre-coaching is less of a concern for Real Talk, as questions are unknown ahead of time, and all responses are required to be recorded synchronously.

As with all online software systems, hardware or internet access issues was often reported by students. While many these reports were likely authentic, in our use of Real Talk we identified a large number of cases where hardware and internet issues were raised as a excuse for not completing a Real Talk discussion. There were also a relatively large number of cases where students did not complete the Real Talk

discussion, though they continued to complete other assessments within Doubtfire. These are potentially cases where the student is looking to deliberately avoid the assessment, and highlights the need for appropriate enforcement of these assessments through hurdle requirements.

7.6. Educator and pedagogical impact

With limitations on the Australian Higher Education sector, rising teacher-student ratios, and rising online-education, educators are often forced to spend less time per student than they would prefer. Further, educators are increasingly concerned with academic misconduct, such as plagiarism and contract cheating. We believe human-in-the-loop, intelligent-teaching assistant systems (Yacef, 2002) described in subsection 2.1 such as Real Talk proposed in this paper can help educators develop trust in online tools, allowing them to focus on more meaningful pedagogical activities such as providing formative feedback, and identifying meaningful learning opportunities.

As discussed, authentication of online student identity can be difficult to implement, with formal invigilated examinations being challenging in the current environment. The system proposed in this paper works silently with the audio tools in the learning and feedback tool to ensure the identity of students engaged in the audio tools are consistent. In cases where the model returns a false negative, the only impact would be that the tutor would manually review the audio examples and determine the likelihood that the case requires further investigation.

In our practice, we ensure each unit or course provides guidelines for when tutors should engage with online students via the audio discussion tools throughout the teaching period, to both foster the benefits of student-tutor audio interaction, and to ensure sufficient examples of student audio is collected for authentication purposes. Renzella and Cain (2020) indicates that the provision of formative audio feedback in programming education environments required less time compared to written feedback in most circumstances, and was well received by experienced tutors.

7.7. Addressing student privacy concerns

The speaker verification approach described in this paper was designed to be privacy-centric and student friendly as we are utilising potentially sensitive student data collected via the learning and feedback tool. For the verification and enrollment purposes, only the MFCC generated from the audio examples is stored. An MFCC representation of speaker audio is not suitable to be reconstructed back into recognisable speech, due to the lossy compression algorithms used to generate the MFCC. While Cloud speaker recognition services[3] are available, and provide simpler mechanisms for integration into existing tools, they require the transmission of student data to potentially costly third-party services. The approach described in this paper can be entirely conducted on institution's internal systems.

7.8. Accessibility considerations

Educational systems should consider the accessibility requirements of their users from the outset:

- **Machine-generated speakers**: Some users utilise technology-assisted speech generators either on a personal computer, or a device to assist their body physiologically produce speech.
- **Recording Anxiety**: A small number of students in our experience with audio discussion tools express anxiety around recording, transmitting and/or storing their voice in such systems.

In these or similar cases it is important educators make a case-by-case decision to avoid the use of systems such as the proposed system in this paper, and to reach out and engage with the student in the best way that suits them.

[3] Such as Microsoft Azure Speaker Recognition Services.

8. Limitations

This specific measurement of the model's performance in adverse or edge-cases was not evaluated as part of this study due to the general scope of the study in evaluating the feasibility of such a system. However, the concern is addressed in the design of the tool as described as a "tutor-in-the-loop" system, wherein cases where the system performs poorly, or misidentifies identities which would cause a false negative (like a lack of sleep made the student fail the identity match), the system's only course of action would only alert the tutor to manually review the case and make a decision.

The dataset used for evaluation consisted of real-world student audio collected over a four-year period in an Australian learning and feedback tool. As such, the dataset was not curated for machine-learning evaluation, and often low-quality, or problematic recordings were included which led to low-scoring or erroneous results from Deep Speaker. To mitigate this, we manually reviewed a large subset of cases to generate the overall accuracy. While this is a valid concern, the dataset is representative of real-world usage which will contain such issues.

A potential threat to the generalisability of this research could be argued from the fact that this project only evaluated Deep Speaker on audio generated from one Australian University's students. While a valid concern, we do not believe this proves a significant threat. Firstly, the Deep Speaker system was designed for both Mandarin and English models. Further, the background and demographic of speakers was sufficiently representative, with a combination of local Australian, and international students included in the evaluation. If the proposed system were to be adopted in a non-English speaking country, then the implementation could be modified to replace the Deep Speaker component with a more suitable model trained on localised speech data.

Detecting contract cheating or student coaching is challenging for educators, and no system including the proposed Real Talk system will be perfect. One way in which students may attempt to circumvent Real Talk would be to receive real-time coaching while they listen to the discussion prompts. Some approaches to combatting this in the future include training educators with strategies for formulating more challenging prompts to coach. Further, exploring the detection of signals such as silent audio (for example, if a student were to mute their microphone during the prompts) as a potential marker of academic misconduct.

While the proposed system provides value in the deterrence of academic misconduct alone, the degree to which the use of the system would provide actionable results in a litigious environment is untested, and is highly dependent on institutional policies.

9. Future work

9.1. Adversarial evaluation of the proposed system

This paper presents an innovation for validating student identity throughout oral assessments. To be implemented in institutions, valuable future work could stress-test the approach. We identify two such areas for adversarial evaluation of the system:

- Solicit third-parties such as contract cheating services, students, or research staff to attempt to circumvent the Real Talk + Deep Speaker system to evaluate the approach. Dawson and Sutherland-Smith (2018) present a similar idea in which the authors solicited contract cheating works to determine if educators could identify cases of contract cheating.
- A study to have a set of speakers attempt to produce audio recordings which "trick" the Deep Speaker model into confirming identity. Such

areas to be explored could be digital voice manipulation, or physiological voice manipulation. It is worth noting however, that in the current design of the system, all discussion recordings are reviewed by educators, so while it is possible that voice manipulation could bypass the validation, the educator would identify the irregularity.

The following items outline future engineering work which could be explored to further strengthen Real Talk:

- Developing support for video-discussions for higher-impact assessment activities.
- Training the Deep Speaker model on more relevant datasets (such as our own student speech corpus).
- Using tutor feedback to reinforce the model over time to improve the accuracy.
- Enhancing the approach with additional markers, such as IP Address detection.

Previous research has explored educator experience and perceptions of Real Talk, however further work is required to capture and evaluated student perceptions and experiences of the tool, especially compared to other proctored assessment tools.

Finally, while Real Talk and the associated systems such as Doubtfire are entirely open-source and available for educators and researchers alike to access, further work is required to make the tools more readily available such as introducing interoperability with prevalent Learning Management Systems.

10. Conclusion

This paper presents an evaluation of a machine learning speaker verification system (Deep Speaker) against a historical speech dataset containing 2,448 examples of student responses to formative feedback and discussion prompts, provided in a learning and feedback tool. The dataset comprised of speech from 33 unit/teaching periods, with Australian and International students of both undergraduate and postgraduate units. The results indicate that such speaker recognition systems can provide a seamless, student-friendly approach for student identity validation.

Real Talk is a student-tutor discussion tool which facilitates student-synchronous, tutor-asynchronous audio-based discussions. The tool ensures that students provide an authentic response to a tutor's discussion prompt by only allowing the student to access the prompt(s) at the moment of response. We present Real Talk as an approach to help educators achieve the benefits of online oral assessment with larger cohorts, not requiring staff to engage in student scheduling activities.

Together, Deep Speaker (or similar speaker verification tools) applied to authentic discussion tools such as Real Talk can provide a secure discussion and assessment opportunity, without requiring manual interaction (such as requiring students to sweep a room with a camera), human invigilation, or proctored assessment tools. The system can alert tutors in instances that the identity of the speaker could not be verified, allowing the tutor to focus their limited time on assessment, providing high-quality formative feedback and identifying learning opportunities.

Acknowledgments

The authors would like to acknowledge Aidan Griffiths, and the undergraduate students who assisted in the development of the Deep Speaker implementation used in this study.

Appendix

A.1) Audio Engine Source Code.

The source code for the audio engine, and Real Talk can be found at the following repository: https://zenodo.org/record/5144543.

A.2) Deep Speaker Implementation.

The source code for the Deep Speaker implementation can be found at the following repository: https://zenodo.org/record/5144543.

References

Akimov, A., & Malin, M. (2020). When old becomes new: A case study of oral examination as an online assessment tool. *Assessment & Evaluation in Higher Education, 45*, 1205–1221. https://doi.org/10.1080/02602938.2020.1730301

Australian Broadcasting Corporation. (2020). *University students at UQ raise concerns about online exam monitoring service ProctorU*. URL: https://www.abc.net.au/news/2020-07-16/uq-privacy-issues-concerns-student-exams/12454964.

Barnes, C., & Paris, B. L. (2013). *An analysis of academic integrity techniques used in online courses at a Southern university*. Technical Report July.

Bretag, T., Harper, R., Burton, M., Ellis, C., Newton, P., Rozenberg, P., Saddiqui, S., & van Haeringen, K. (2019). Contract cheating: A survey of Australian university students. *Studies in Higher Education, 44*, 1837–1856. https://doi.org/10.1080/03075079.2018.1462788

Clarke, R., & Lancaster, T. (2006). Eliminating the successor to plagiarism? Identifying the usage of contract cheating sites. Gateshead, UK. In *Proceedings of the 2nd international plagiarism conference*. URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.120.5440&rep=rep1&type=pdf.

Cummaudo, A., Barnett, S., Vasa, R., & Grundy, J. (2020). Threshy: Supporting safe usage of intelligent web services. In *ESEC/FSE 2020 - proceedings of the 28th ACM joint meeting European software engineering conference and symposium on the foundations of software engineering* (pp. 1645–1649). Association for Computing Machinery, Inc. https://doi.org/10.1145/3368089.3417919.arXiv:2008.08252.

Dawson, P., & Sutherland-Smith, W. (2018). Can markers detect contract cheating? Results from a pilot study. *Assessment & Evaluation in Higher Education, 43*, 286–293. https://doi.org/10.1080/02602938.2017.1336746

Department of Education and Training. (2018). *uCube - higher education statistics*. Technical Report. URL: http://highereducationstatistics.education.gov.au/.

Evans, J. (2018). *15% of students admit to buying essays. What can universities do about it?*. URL: https://theconversation.com/15-of-students-admit-to-buying-essays-what-can-universities-do-about-it-103101.

García-Peñalvo, F. J., Corell, A., Abella-García, V., & Grande-de Prado, M. (2021). Recommendations for mandatory online assessment in higher education during the COVID-19 pandemic. In *Lecture notes in educational technology* (pp. 85–98). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-981-15-7869-4_6.

Gleaves, A., & Walker, C. (2013). Richness, redundancy or relational salience? A comparison of the effect of textual and aural feedback modes on knowledge elaboration in higher education students' work. *Computers & Education, 62*, 249–261. https://doi.org/10.1016/j.compedu.2012.11.004

Google Inc. (2019). *WebRTC home*. URL: https://webrtc.org/.

Harper, R., Bretag, T., Ellis, C., Newton, P., Rozenberg, P., Saddiqui, S., & van Haeringen, K. (2019). Contract cheating: A survey of Australian university staff. *Studies in Higher Education, 44*, 1857–1873. https://doi.org/10.1080/03075079.2018.1462789

Harper, R., Bretag, T., & Rundle, K. (2021). Detecting contract cheating: Examining the role of assessment type. *Higher Education Research and Development, 40*, 263–278. https://doi.org/10.1080/07294360.2020.1724899

Haskell-Dowland, P. (2020). *ANU will invigilate exams using remote software, and many students are unhappy*. The Conversation. URL: https://theconversation.com/anu-will-invigilate-exams-using-remote-software-and-many-students-are-unhappy-137067.

Joughin, G. (1998). Dimensions of oral assessment. *Assessment & Evaluation in Higher Education, 23*, 367–378. https://doi.org/10.1080/0260293980230404

Kumar, R. (2020). *Assessing higher education in COVID-19 Era*. Technical Report 2. URL:, 10.26522/brocked.v29i2.841 https://journals.library.brocku.ca/brocked.

Lancaster, T., & Clarke, R. (2016). *Contract cheating: The outsourcing of assessed student work*. Handbook of Academic Integrity. https://doi.org/10.1007/978-981-287-098-8_17

Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., Cao, Y., Kannan, A., & Zhu, Z. (2017). *Deep speaker: An end-to-end neural speaker embedding system*. arXiv:1705.02304.

Merry, S., & Orsmond, P. (2008). Students' attitudes to and usage of academic feedback provided via audio files. *Bioscience Education, 11*, 1–11. https://doi.org/10.3108/beej.11.3

Newton, P. M. (2018). How common is commercial contract cheating in higher education and is it increasing? *Systematic Reviews*. https://doi.org/10.3389/feduc.2018.00067

O'reilly, G., & Creagh, J. (2016). A categorization of online proctoring. *Global Learn*, 542–552, 2016.

Palvia, S., Aeron, P., Gupta, P., Mahapatra, D., Parida, R., Rosner, R., & Sindhi, S. (2018). *Online education: Worldwide status, challenges, trends, and implications*. https://doi.org/10.1080/1097198X.2018.1542262

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *ICASSP, IEEE international conference on acoustics, speech and signal processing - proceedings* (pp. 5206–5210). Institute of Electrical and Electronics Engineers Inc. volume. https://doi.org/10.1109/ICASSP.2015.7178964, 2015-Augus.

Parkes, M., & Fletcher, P. (2017). A longitudinal, quantitative study of student attitudes towards audio feedback for assessment. *Assessment & Evaluation in Higher Education, 42*, 1046–1053. https://doi.org/10.1080/02602938.2016.1224810

Rémy, P. (2020). *Deep speaker: An end-to-end neural speaker embedding system*. URL: https://github.com/philipperemy/deep-speaker.

Renzella, J., & Cain, A. (2017). Supporting better formative feedback in task-oriented portfolio assessment. TALE 2017. In *Proceedings of 2017 IEEE international conference on teaching, assessment and learning for engineering* (pp. 360–367). https://doi.org/10.1109/TALE.2017.8252362, 2018-Janua.

Renzella, J., & Cain, A. (2020). Enriching programming student feedback with audio comments. *Proceedings - International Conference on Software Engineering*, 173–183. https://doi.org/10.1145/3377814.3381712

Renzella, J., Cain, A., & Schneider, J. G. (2021a). Real Talk: Illuminating online student understanding with authentic discussion tools. In *SIGCSE 2021 - proceedings of the 52nd ACM technical symposium on computer science education* (pp. 886–892). Association for Computing Machinery, Inc. https://doi.org/10.1145/3408877.3432484.

Renzella, J., & Griffiths, A. (2021). *Deep Speaker-based project source code*. URL:, 10.5281/zenodo.5144543 https://zenodo.org/record/5144543.

Renzella, J., Tubino, L., Cain, A., & Schneider, J.-G. (2021b). Enhancing online education with intelligent discussion tools. COVID-19 and Education: Learning and Teaching in a Pandemic-Constrained Environment. In T. Luo (Ed.), *Anthony scime, christopher cheong, jo coldwell-neilson, kathryn MacCallum* (pp. 75–97). Informing Science Press.

Rowland, S., Slade, C., Wong, K. S., & Whiting, B. (2018). 'Just turn to us': The persuasive features of contract cheating websites. *Assessment & Evaluation in Higher Education, 43*, 652–665. https://doi.org/10.1080/02602938.2017.1391948

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology, 67*, 415–437. https://doi.org/10.1146/annurev-psych-073115-103235

Velliaris, D. M. (2016). *Handbook of research on academic misconduct in higher education* (pp. 1–439). Handbook of Research on Academic Misconduct in Higher Education. https://doi.org/10.4018/978-1-5225-1610-1

Vlachopoulos, D. (2020). Covid-19: Threat or opportunity for online education? *Higher Learning Research Communications, 10*. https://doi.org/10.18870/hlrc.v10i1.1179

Walker, M., & Townley, C. (2012). Contract cheating: A new challenge for academic honesty? *Journal of Academic Ethics, 10*, 27–44. https://doi.org/10.1007/s10805-012-9150-y

Wood, K. A., Moskovitz, C., & Valiga, T. M. (2011). Audio feedback for student writing in online nursing courses: Exploring student and instructor reactions. *Journal of Nursing Education, 50*, 540–543. https://doi.org/10.3928/01484834-20110616-04

Yacef, K. (2002). Intelligent teaching assistant systems, 136–140. In *Proceedings - international conference on computers in education, ICCE 2002*. https://doi.org/10.1109/CIE.2002.1185885. IEEE Comput. Soc volume 1 of *International Conference on Computers in Education*.