Automated Speech Recognition in language learning: Potential models, benefits and impact

by Michael Carrier

Michael Carrier Highdale Consulting michael@highdale.org

Published in Training, Language and Culture Vol 1 Issue 1 (2017) pp. 46-61 doi: 10.29366/2017tlc.1.1.3

Recommended citation format: Carrier, M. (2017). Automated Speech Recognition in language learning: Potential models, benefits and impact. *Training, Language and Culture, 1*(1), 46-61. doi: 10.29366/2017tlc.1.1.3

The study considers Automated Speech Recognition (ASR) in language learning arguing that speech recognition has reached a level of accuracy where it is powering automatic translation and testing. The author considers the impact of ASR technology on language teaching, describes the process of the development of appropriate pedagogical models, and explains how to prepare teachers for their application. The study will give a critical analysis of the pedagogical uses and dangers of ASR technology and address how ASR can be used to automate language assessment.

KEYWORDS: *Automated Speech Recognition, ASR, ELT, speech-to-speech translation, translation software, speech synthesis, automated assessment*

1. INTRODUCTION

The technology of Automated Speech Recognition (ASR) is rapidly becoming more sophisticated and is becoming part of everyday life. The aim of this paper is to outline the impact on English language teaching of the use of automated speech recognition (ASR) technology. I will discuss the nature of ASR, how it works and how it can be used in class.

I will also touch on the technology of speech to speech translation, which uses ASR and speech synthesis, and outline how this may also impact on student motivation and teacher course design. Finally, I will address how ASR can be used to automate language assessment.

2. THEORETICAL BACKGROUND

2.1 What is ASR?

Automated Speech Recognition (ASR) converts audio streams of speech into written text. ASR is still imperfect but improving rapidly in terms of its accuracy in recognising spoken discourse and transcribing it into written text.

ASR is based on big data-searching language corpora and finding matching patterns in data in order to convert the audio into written text. However, it does not analyse the audio semantically. The ASR output cannot assess meaning or coherence – it is not the same as Natural Language Processing which parses and analyses language. It merely transcribes speech

'ASR is based on big data-searching language corpora and finding matching patterns in data in order to convert the audio into written text'

and turns spoken language into written language – using complex statistical and language analysis models.

ASR's rate of accuracy is increasing and it is now being used extensively in commercial applications and in telecommunications. There are two main types of commercial ASR application – small vocabulary/many-users vs large vocabulary/limited-users.

Small vocabulary/many-users is ideal for automated telephone answering. The users can speak with a great deal of variation in accent and speech patterns, and the system will still understand them most of the time. However, usage is limited to a small number of predetermined commands and inputs, such as basic menu options or numbers.

Large vocabulary/limited-users systems work with a good degree of accuracy (85% or higher with an expert user) and have vocabularies in the tens of thousands of words. However, they may initially need training with known texts to work best with a small number of primary users. Earlier systems have a lower accuracy rate with an unknown or 'untrained' user but this is changing rapidly. The former system is not of much relevance to education, but the latter system can now be applied to educational use, where its application to language learning is extremely relevant to modern digital learning practices.

2.2 ASR and English language teaching (ELT)

ASR has a chequered history in ELT, and there have been many inadequate commercial English language learning software products which promised more than they delivered. Earlier computer-assisted language learning (CALL) systems, sometimes called 'Voice-interactive CALL', attempted to use ASR for pronunciation training or to initiate dialogic interactions.

In most cases until recently, the low level of ASR accuracy meant that this became a frustrating and unsatisfactory learning experience for students as there were too many false positives or false negatives in the processing of their speech. This gave ASR in ELT a poor reputation but with new technological developments in the last few years leading to new levels of accuracy and recognition quality new opportunities arise. ASR can facilitate new ways to work on phonology and accent – various applications are able to 'listen' to a learner's pronunciation and provide formative assessment and feedback on the accuracy of the

doi: 10.29366/2017tlc.1.1.3
rudn.tlcjournal.org

'Speech-enabled translation software and apps allow students to speak into a phone or tablet and instantly hear the spoken translation in the target language'

pronunciation. ASR can also facilitate responses to communicative interactions in the classroom, where students can use their tablets (in pairs) to speak or write responses to a task and get instant correction or formative assessment of their pronunciation or comprehensibility.

In addition, ASR can facilitate automatic translation. Speech-enabled translation software and apps allow students to speak into a phone or tablet and instantly hear the spoken translation in the target language (L2). ASR can also facilitate computer-based automated marking of ELT examinations – both written and spoken exams, with an accuracy approaching that of human assessors.

2.3 How does ASR work?

ASR turns speech into written text by using a 'speech recognition engine'. Speech recognition engines are software systems that take the audio output produced by users' speech (i.e. digital audio signals from a microphone or recording) and process the stream of sound into individual sounds, vowels and consonants, lexical items and outputs a written transcription of the speech. Speech recognition engines need the following components:

- an acoustic model which takes audio recordings of speech and their transcriptions, and 'compiles' them into a statistical representation of the sounds that make up each word;
- a language model is a data file containing the probabilities of sequences of words. A grammar file is a much smaller data file containing sets of predefined combinations of words.

2.4 How can we use ASR in language education?

2.4.1 Pedagogical approaches

We can benefit from speech recognition to support many aspects of teaching, learning and assessment in language education. ASR can be used for a variety of classroom and out-of-class activities, such as dictation, voice search (e.g. via Google), pronunciation practice, vocabulary and grammar exercises, translation practice, and marking of student production.

ASR can, for example, facilitate new ways to work on phonology and accent, using computer-aided pronunciation teaching (CAPT) software to listen to and give feedback on learners' pronunciation.

One of the best-known examples of this was the

'We can benefit from speech recognition to support many aspects of teaching, learning and assessment in language education'

IBM programme *Reading Companion*, which could listen to the student's pronunciation of a known text (usually a short story text) and correct the pronunciation it heard, providing helpful feedback. It achieved this by comparing the learner's pronunciation (post-ASR) to the stored model pronunciations it expected to hear. IBM provided this speech-enabled literacy skill development free for students around the world, especially in developing countries with PC labs available. ASR can also facilitate responses to communicative interactions in the classroom, where students can use their tablets (in pairs) to speak or write responses to a task, and get an instant correction or formative assessment.

Very little research seems to have been carried out on designing pedagogically-appropriate activities to take advantage of this new technology. Escudero-Mancebo et al. (2015) report on the use of 'minimal pair' games to develop pronunciation skills using ASR applications on Android devices (smartphones and tablets). As in many CAPT applications, however, the interaction was essentially self-study: between learner and device, with no learner-to-learner interaction. This form of pedagogical model may be useful in skills development and assessing pronunciation, but it is to be hoped that further research will be done on pair and groupwork-based activities such as those suggested below, to ensure the classroom remains a communicative space.

2.4.2 Examples of available applications

A smartphone app called *Speaking Pal* provides activities where the learner interacts with a virtual character, shown in short video clips, and the character responds. *Speaking Pal* suggests this pedagogical approach with specific reference to learning English:

- mini-lessons enable students to learn English in small sections that last 5 minutes or less (micro-learning);
- this unique methodology allows students to learn effortlessly during their daily activities.

Speaking Pal contains a programme called *English Tutor* with short, real-life dialogues where the user controls the conversation flow, like in a real video or phone call. *English Tutor* is able to provide instant feedback on the student's speaking performance, assuming that the ASR engine actually recognises what the student says.

A review from ELTjam (Gifford, 2013) suggests that this is not always successful.

doi: 10.29366/2017tlc.1.1.3
rudn.tlcjournal.org

'The 'feedback' consisted of showing me which words I'd said 'well' (in green) and which ones I needed to work on (red). There doesn't seem to be an explanation as to what in particular the problem is with my speaking, so I'm none the wiser as to how to improve. I just tried to shout it a couple of times to be sure' (Gifford, 2013).

Velawoods English, a self-study course from Velawoods and Cambridge University Press, uses speech recognition from Carnegie Speech to give students feedback on their utterances and pronunciation. It provides a game-like environment where the learner interacts with characters, as with *SpeakingPal*. In this case the course is more structured and language presented is practised in conversations with the characters – with learners' responses checked by speech recognition.

2.4.3 Self-study activities

Both of the examples described above tend to underpin criticisms of ASR-based activities, namely that this leads to students speaking by themselves in isolation. However, for many learners, self-study activities are the only way to gain enough language learning time and language exposure to make progress, either because they cannot attend class or their class hours are limited (e.g. in secondary schools).

As noted above, ASR can be used to get diagnostic feedback on pronunciation issues using CAPT activities, and for many students this alone would be a welcome support to their learning, building spoken confidence when they see their correct pronunciation is recognised by the ASR engine.

Speech recognition can help learners to engage in speaking, even without a partner being available. The student works alone to dictate a text or act out a dialogue given by the teacher, dictating to a device. The ASR system provides written output, which the student can then review and correct and share with the teacher for review.

2.4.4 Classroom/communicative activities

ASR can be used for collaborative activities such as storytelling. Students tell a story by dictating to their device. A pair or group design a story, and one student takes the dictating role to ensure user accuracy. Other students edit the resulting text and check accuracy and appropriacy using online dictionaries.

ASR can be used for Conversation tasks where, for example, students in pairs write a dialogue and perform it as dictation, then read and correct the written output. The process of reading the written output of the ASR engine provides students with a feedback loop on pronunciation and speech structure, allowing them to reflect on what they said; if the ASR engine recognises what they said, then it is most likely pronounced correctly, but

students still need to analyse whether it is grammatically correct and appropriate in discourse terms. If it is not recognised, then what was said was most likely pronounced incorrectly and this can be improved and modified.

If the written output is reviewed by a spellchecker and grammar checker (as in Microsoft Word) students can get further formative feedback on lexis and grammar choices, and can improve their performance. Students in pairs can use Google Voice Search to discover information for a writing task – finding information about the history of their city or country, for example – and check that their spoken performance returns the expected search results.

Similarly, student pairs/groups can use the digital assistants they have access to (e.g. Siri on phones or Alexa devices) to practice asking for and processing information in real-life scenarios.

2.5 Speech synthesis

This paper has touched on the developments in automatic speech recognition. But the recognition of speech is of course only one side of the story. Computers can listen to, understand and use what we are saying, at least to a certain extent (e.g. for translation purposes) – but can they reply?

The next stage of technological development that we can see emerging is helping computers to produce speech, to enable a dialogue between learner and computer, where the computer responds to spoken input from the student. In the science fiction context we expect humans and computers to be able to converse. We are not yet at this stage, so it is unrealistic to expect that humans and computers can have naturalistic conversations or that language learners can speak to a computer and expect a detailed reply.

'Speech recognition can help learners to engage in speaking, even without a partner being available'

However, strides are being made in this direction, and it may be that computer speech is sufficiently well developed for some simple language learning scenarios to be used in this way. Computers have long been able to create speech, using 'speech synthesis' software otherwise known as text-to-speech (TTS) engines. This uses the basic audio features of a computer to create spoken output from written text. In one model of text-to-speech production, the computer uses recordings of very short segments of human speech, which can be combined to create sentences. This can be heard typically in transport announcements, for example. A more powerful model is when the computer creates sound waves that mimic human speech, using the same audio systems that allow it to play music. This is more flexible, as any text can be

doi: 10.29366/2017tlc.1.1.3
rudn.tlcjournal.org

'Early versions of this have been justly derided as sounding artificial and robotic, and many people are familiar with the sort of speech synthesis generated for people like Stephen Hawking'

produced as speech, but it is more difficult to make the speech sound natural. Early versions of this have been justly derided as sounding artificial and robotic, and many people are familiar with the sort of speech synthesis generated for people like Stephen Hawking.

The technological problem is not in the creation of natural speech, in the sense of sounds that we can understand as spoken English, but in the creation of what to say.

How does the computer know what to talk about, which words to pick from its vocabulary database, to arrange in a certain order and to give a natural pronunciation and intonation output? For several years there have been language learning software programs and apps that attempt to create a dialogic experience, where the learner speaks to the device and receives spoken feedback. In most cases this is still at a very primitive stage and it will most likely be several years more before this can be developed fully (cf. *Speaking Pal* example above).

We can be fairly certain that language teachers will not be made redundant in the near future, but the motivations and aspirations of their learners, and the methodologies and technologies used in the teaching process, may become significantly different.

2.6 Speech-to-speech translation

Speech-to-speech translation seems like science fiction. It first surfaced as a popular concept in Adams' (1979) book, *The Hitchhiker's Guide to the Galaxy*, where he described the Babel Fish, a small insert you put in your ear which allowed aliens to speak to each other across the universe by providing instant translations. This is now a reality (including the ear-based unit). Speech-to-speech translation, also referred to as speech-enabled translation (SET), means simply that you can speak to a device in your own language and it will automatically 'understand' you, translate the message into another language, and immediately speak that message out loud to the conversation partner, in their own language. It is used in currently available products and services like *Google Translate, Apple Watch* and *Skype Translate*. This approach combines the technology of speech recognition as described above with machine translation models. The ASR process outputs the L1 speech as text, which can be processed by the translation engine. This uses a statistically-based machine translation system, which essentially looks for patterns in hundreds of

millions of documents to help decide on the best translation, using a corpus of documents that have already been translated between the pair of languages requested, originally by human translators.

Finding patterns in this corpus helps the machine translation to make a statistically-informed intelligent guess about the best translation into L2. *Microsoft Translator* is a linguistically informed statistical machine translation service built on more than a decade of natural-language research at Microsoft. Rather than writing hand-crafted rules to translate between languages, modern translation systems approach translation as a problem of learning the transformation of text between languages from existing human translations and leveraging recent advances in applied statistics and machine learning. Of course, the resulting translation cannot be as accurate as a human translator, as it is looking at patterns in documents, not analysing and understanding meaning. But it can often provide a useful first translation, and the degree of accuracy is improving rapidly. This L2 machine translation output, as text, is fed into a speech synthesiser, which speaks the result out loud.

3. STUDY AND RESULTS

3.1 Using speech-to-speech translation in class

As noted above, there seems to be very little research carried out on designing pedagogically

'Of course, the resulting translation cannot be as accurate as a human translator, as it is looking at patterns in documents, not analysing and understanding meaning'

appropriate activities for the use of speech-enabled translation (SET). This seems quite surprising, as the relevance of a speech translator for L1 to L2 to a learner who is trying to become proficient at speaking the L2 language in order to be understood by L2 speakers, would seem to be quite significant.

There is also a 'real world' issue – people are already using text-based and speech – enabled translation on their phones. it is becoming part of everyday experience, and so it seems unusual not to recognise this in classroom practice. Should we make space for it in our pedagogical approach?

There is as yet no established pedagogical model for using speech-enabled translation in the classroom, but we can suggest that a similar approach is taken to that suggested above for ASR activities in both self-study and communicative modes.

The core pedagogical principle is of using ASR and SET to give learners the opportunity of reflection

doi: 10.29366/2017tlc.1.1.3
rudn.tlcjournal.org

on the quality of their spoken performance by getting ASR and SET feedback. In addition, the production and monitoring of this spoken performance can be managed by a communicative and cooperative activity – utilising learners' creativity in dialogue construction and storytelling, for example, rather than restricting learners to what Thornbury (2005) calls 'vocalising grammar': *'Speaking activities are often simply exercises in vocalising grammar, as if this were all that were needed'* (Thornbury, 2005, p. 11).

The core process could be seen as:

- learn relevant language input
- discuss task
- speak/record in pair work
- check meaning via SET translation
- discuss differences in group/with teacher
- post-task reflective writing exercise

Bouillon et al. (2011) describe a pedagogical activity based on translation. The learner gets a prompt in L1 and is asked to speak a word or phrase in L2. The system recognises the speech in L2 and translates it back to L1, so that the learner can see if it matches the prompt.

The interesting pedagogical issue here is that it does not ask for a translation of the prompt, but the prompt is an instruction like *'Book a table for 1 person'* or *'Say your age',* thus avoiding simplistic L1/L2 correlations in the learner's mind. The ASR-based activity opens up: *'A powerful and interesting perspective, because different types of new cognitive processes are now involved: i) translation and transformation of a rich linguistic input (gloss) from L1 to L2; ii) speaking aloud and being checked by speech recognition; iii) reading what has been recognised (correctly or incorrectly); iv) listening and reading the correct answers (learner's production vs. the mother tongue version)'* (Bouillon et al., 2011, p. 269).

Further to the suggested activities, some possible activities for speech-enabled translation could include the following.

Activity 1. A simple activity is for students to select or write a text in their L1, that they will then translate into L2 (e.g. English). The task is:

- Using Google Translate, students write a sentence or short text in L1.
- Student A translates it into L2 (English) in writing.
- Student B then speaks the L2 (English) text into Google Translate and hears it translated back into L1, and spoken by the device.

The two students then compare the versions of the text, and note differences created by the translation, asking for teacher guidance where needed. The translation direction can also be reversed to see where discrepancies arise.

Activity 2. Access to communicative practice is often problematic for students, so SET may help to create an opportunity (albeit slightly artificial) to provide some conversation practice.

Two students decide on a conversation topic, where student 1 speaks in English, for example, and uses Google Translate or another SET to create an L1 translation, which student 2 responds to in L1, creating an L2 translation using SET. The students carry on a conversation using Google Translate but making notes of the nature of the translation choices made by the programme, to show discrepancies or to discuss alternative translations. Variations on this involve both students speaking only in L2 but using the SET to check their utterances in their L1 translation. Lower level students can speak in L1 and get the translation in L2 to use as semi-authentic comprehension input.

3.2 Impact of speech-to-speech translation on learner motivation

As the quality of translation improves, it is important to remember the restrictions that the process operates under – the translation engine does not actually understand the meaning of the text being translated, so errors will always occur – though they may become more minor or unusual. Computers have got much better at translation, voice recognition and speech synthesis, but they still don't understand the meaning of language.

The statistical approach, using 'big data', can only take translation accuracy so far: as with the translation model the language model uses a brute force statistical approach to learn from the training data, then ranks the outputs from the translation model in order of plausibility.

Further improvements are expected from the use of deep learning, based on the use of digital neural networks. This mimics the way that neurons in the brain hold information and link to each other, and a computerised neural net can be used to create translations. The neural net is trained on large amounts of data, which consist of texts in one language and their translation in other.

The neural net 'learns' from the contexts of these existing translations and is able then to provide an approximate translation of any new text given to it, using statistical analysis rather than analysing the language input using rule based systems.

Will people be more or less likely to wish to learn a language to a high level of proficiency for their professional use – reading professional literature, attending conferences, corresponding with clients – if two-way translation of text and automatic speech-to-speech translation in live conversations can be carried out by computers?

This is an area where further research into affective issues is required. If you can send emails and carry

doi: 10.29366/2017tlc.1.1.3
rudn.tlcjournal.org

'Further improvements are expected from the use of deep learning, based on the use of digital neural networks'

out telephone conversations with speakers of other languages without learning the language itself, this might lead to a lack of motivation. Conversely, the fact that you can build contacts with people from another language community may provide the affective motivation to get further involved with that language community. The sense of success, that talking to somebody from another language group is possible and enjoyable and useful, may be the trigger that makes people decide to learn at least the basics of the language for themselves (Ye & Young, 2005).

By the same token, the quality deficiencies in machine translation and speech to speech translation may spur some learners to greater levels of proficiency so that they can 'master' the automatic translation systems and improve on them.

3.3 Automated assessment

3.3.1 Autimating marking and grading

Marking students' speech output, marking tests (whether written or oral) and giving individual feedback is one of the most time-consuming tasks that a teacher can face. Automating the process makes sense, argues Dr Nick Saville, Director of Research and Validation at Cambridge Assessment:

'Humans are good teachers because they show understanding of people's problems, but machines are good at dealing with routine things and large amounts of data, seeing patterns, and giving feedback that the teacher or the learner can use. These tools can free up the teacher's time to focus on actual teaching' (Saville, 2015).

ASR technology can be used to automate the marking and grading of students' spoken production (as well as written production, see below). Speech produced in assessment situations can be fed through the same speech recognition process as described above, and the written text output produced then analysed and graded by being compared to a corpus of spoken language produced by students at different levels of the Common European Framework of Reference (CEFR).

This can provide the benefit of speed – the spoken output from students does not need to be recorded and sent to a human examiner, but can be graded instantly, as Saville explains.

'Automated assessment won't replace human examiners anytime soon, but it can add great value to their work. For example, it can provide additional layers of quality control, speed up

processes and allow teachers to offer more objective in-course tests which give detailed diagnostic feedback to help students to improve their English more effectively' (Saville, 2015).

3.3.2 How does it work?

The technical details of automated assessment of speech are beyond the scope of this paper, but the core process is quite straightforward and educators need to be aware of the basic approach so that they can feel confident in the results. Student speech is analysed by an ASR engine as in earlier applications above, and turned into transcribed text.

The speech is analysed both in terms of its content (i.e. what is in the transcribed text) and its pronunciation and prosody features (i.e. what it sounds like in the audio version of the speech). The speech is analysed by various assessment engines (called 'feature extraction' and 'classifier', for example). The analysis can assess how close the pronunciation is to that of a native speaker, whether there is L1 interference (where the L1 is known) and whether L2 phones are missing from the speech.

The features that can be extracted include the speaking rate (in words per second), the duration of phones (i.e. English sounds) and the length of silences between words; and 'scores' to reflect the closeness to native speaker production (e.g. more or less L1 interference) and acoustic precision.

'What the analysis cannot currently do is to assess the appropriacy and coherence of the speech – is this an appropriate response to the question asked in the test?'

What the analysis cannot currently do is to assess the appropriacy and coherence of the speech – is this an appropriate response to the question asked in the test? Is this content, whether grammatically and lexically correct or not, appropriate to the topic of the task? Current research uses very complex algorithms to assess topic relevance, with increasingly positive results.

Thus, the automated system can be relatively accurate in evaluating pronunciation and comparing it to L1 speaker models, and evaluating fluency (hesitations, pauses, speed, partial words, etc.) but cannot evaluate the meaning and the coherence of the topic discussed.

Within the bounds of these constraints, this automated analysis is still able to produce an estimated CEFR grade (in terms of CEFR levels, A1, A2, etc.). The accuracy of this grade is measured by comparing the grades created by automated machine-rating with the grades produced by

doi: 10.29366/2017tlc.1.1.3
rudn.tlcjournal.org

human assessors who have listened to the same spoken production. In work done by the ALTA Institute (Litman et al., 2016), the correlation is now in the region of 0.87, meaning the automated grading system reaches the same grade as the human rate in 87% of the cases.

Advances in speech recognition and machine learning mean that computers can now complement the work of human assessors, giving surprisingly accurate evaluations of language and helping to diagnose areas for improvement. This may seem surprising, as speech is such a complex area to grade, even for human assessors. Future developments are likely to increase this correlation to close to 100%. Even given these constraints, automated assessment of speech can still be extremely useful in supporting teachers in three key areas: diagnostic evaluation, feedback loops for learners, and low stakes practice assessments. This approach could be brought into play especially in large-scale education ministry projects where large numbers of learners (or teachers) need to be evaluated across a wide geographical area. A low stakes automated speech assessment would be much easier, and much less expensive, to carry out than flying human assessors around the country.

3.3.3 Automated marking of writing

There is also substantial research in the area of automated assessment of student writing, using similar technologies but operating on text rather than speech. It is beyond the scope of this paper to go into details, but a brief example of learner-oriented experiments in this area can be found at the *Write & Improve* project. This is an online learning system, or 'computer tutor', to help English language learners enhance their writing output by receiving formative assessment from an automatic computer system.

Write & Improve is built on information from almost 65 million words gathered over a 20-year period from Cambridge English tests taken by real exam candidates, who were speaking 148 different languages and living in 217 different countries or territories. Built by Professor Ted Briscoe's team in Cambridge University's Computer Laboratory, it is an example of a new kind of tool that uses natural language processing and machine learning to assess and give guidance on text it has never seen before, and to do this indistinguishably from a human examiner.

3.3.4. Impact on assessment and the classroom

If automated assessment continues to improve in technology and accuracy, what impact could it have on the teaching and learning of languages? There is an obvious impact on testing and assessment systems, but what additional impact would there be on classroom pedagogy, teaching methodology, and learning activities, because of the 'washback' effect of this kind of assessment? It

'Advances in speech recognition and machine learning mean that computers can now complement the work of human assessors, giving surprisingly accurate evaluations of language and helping to diagnose areas for improvement'

seems likely that the availability of easily administered automated assessments would lead to more frequent use of assessment tools – whether in the classroom or outside.

The use of automated assessment would allow regular formative assessment to be carried out, with the provision of feedback to learners and teachers, without increasing the teachers' already heavy workload or reducing classroom contact time (Yuan & Briscoe, 2016). Aspects of this development can already be seen in the inclusion of adaptive testing tools and 'Learning Oriented Assessment' (LOA) methodologies into new coursebooks. Further research into the impact of this on learning outcomes is urgently needed.

4. CONCLUSION

Human beings are speech-driven, and anything which allows them to carry out daily tasks using smartphones and computers without unnecessary typing will be extremely popular, both in commercial and educational applications. It is clear that the growing use of portable and wearable devices such as watches and even glasses will lead to more ubiquitous use of speech recognition.

Whatever can be speech-enabled, will become speech – enabled to bring greater speed and convenience to everyday and educational tasks. One can predict the emergence of some future applications of ASR, which only a few years ago would have been considered science fiction, and yet are now being rolled out as products.

Speech to printed output. People still rely upon written documents, and it will soon be possible for people to dictate a letter, language activity response or creative composition simply by speaking out loud to a speech-enabled phone or PC plus wireless printer. This will mean learners and other users can have their learning materials, or content they have created, immediately printed as a hard copy by a standard computer printer situated nearby, without any intermediate typing. This will help students with lower literacy and keyboard skills, as well as disabled students who find typing difficult.

Speech-activated equipment. People have to interact with a wide range of equipment and devices, switching them on and off, changing TV

doi: 10.29366/2017tlc.1.1.3
rudn.tlcjournal.org

channels, choosing washing machine temperatures, dishwasher cycles, thermostat settings et cetera. All of this will be speech-enabled in the very near future and machines will respond to spoken commands. Indeed, some thermostats are already on the market which allow users to choose a new temperature by voice, and some televisions are already being sold with this feature installed. This will impact the classroom as well, though the proliferation of different voices in a large group may mean that the feature often needs to be disabled.

Speechprint 'Star Trek' ID systems. The development of the technology will soon lead to the realisation of every Star Trek fan's dream, the ability to open doors by speaking to them and thereby identifying oneself by a speech print or voiceprint. This is already being trialled as a way to replace passwords for logging into computer accounts.

More widespread automatic marking of speech. The grading of speech is likely to continue to develop very rapidly. For many less advantaged students, with lower literacy and writing skills, speaking is easier than writing an exam paper, and assessment carried out by automated speech recognition may well be fairer to these types of student than traditional examinations would be. Speech is the future. Speech recognition in education, and specifically speech-enabled language education, is definitely here to stay. Our major challenge is to ensure that we are able to update pedagogical approaches and materials to best utilise this technology, and to support teachers in developing the digital skills they need to manage and benefit from this technology.

References

Adams, D. (1979). *Hitchhiker's guide to the galaxy*. London, UK: Pan Books.

Bouillon, P., Cervini, C., Mandich, A., Rayner, E., & Tsourakis, N. (2011, October). Speech recognition for online language learning: Connecting CALL-SLT and DALIA. In *Proceedings of the International Conference on ICT for Language Learning* (pp. 269-274). Florence, Italy: Simonelli Editore.

Escudero-Mancebo, D., Camara-Arenas, E., Tejedor-Garcia, C., Gonzalez-Ferreras, C., & Cardenosos-Payo, V. (2015, September). Implementation and test of a serious game based on minimal pairs for pronunciation training. *In Proceedings of SLaTE 2015: Workshop on Speech and Language Technology in Education* (pp. 125-130). Leipzig: ISCA Special Interest Group SLaTE.

Litman, D., Young, S., Gales, M., Knill, K., Ottewell, K., van Dalen, R., & Vandyke, D. (2016, August). Towards using conversations with spoken

dialogue systems in the automated assessment of non-native speakers of English. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 270-275). Saarbrücken, Germany: SIGdial.

Gifford, T. (2013, November). Speech Recognition apps for ELT: SpeakingPal English tutor. *ELTjam*. Retrieved from https://eltjam.com/speech-recognition-apps-for-elt-speakingpal-english-tutor

Saville, N. (2015, June). Computer tutor. *University of Cambridge*. Retrieved from http://www.cam.ac.uk/research/features/computer-tutor#sthash.NmmbdTSh.dpuf

Thornbury, S. (2005). Awareness, appropriation and autonomy. *English Language Teaching Professional, 40,* 11-13.

Ye, H., & Young, S. J. (2005, September). Improving the speech recognition performance of beginners in spoken conversational interaction for language learning. In *Proceedings of INTERSPEECH2005: 9th European Conference on Speech Communication and Technology* (pp. 289-292). Lisbon, Portugal: ISCA.

Yuan, Z., & Briscoe, T. (2016). Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 380-386). San Diego, CA: ASL.