JOURNAL OF INTERACTIVE MEDIA IN EDUCATION

Hillaire, G, et al. 2019. Humanising Text-to-Speech Through Emotional Expression in Online Courses. *Journal of Interactive Media in Education,* 2019(1): 12, pp. 1–9. DOI: https://doi.org/10.5334/jime.519

ARTICLE

Humanising Text-to-Speech Through Emotional Expression in Online Courses

Garron Hillaire, Francisco Iniesto and Bart Rienties

This paper outlines an innovative approach to evaluating the emotional content of three online courses using the affective computing approach of prosody detection on two different text-to-speech (TTS) voices in conjunction with human raters judging the emotional content of the text. This work intends to establish the potential variation on the emotional delivery of online educational resources through the use of a synthetic voice, which automatically articulates text into audio. Preliminary results from this pilot research suggest that about one out of every three sentences (35%) in a Massive Open Online Course (MOOC) contained emotional text and two existing assistive technology voices had poor emotional alignment when reading this text. Synthetic voices were more likely to be overly negative when considering their expression as compared to the emotional content of the text they are reading, which was most frequently neutral. We also analysed a synthetic voice for which we configured the emotional expression to align with course text, which showed promising improvements.

Keywords: emotions; accessibility; MOOCs; Online Learning; text-to-speech

Introduction

In the field of online education in higher education, there are various challenges for learners on how to make learning material accessible. For example, one of the most frequently mentioned barriers is that learners perceive a lack of resources to participate in various online educational activities (Miles, 2000). In many countries, legal protections ensure considerations for learners with disabilities when providing online education. However, in practice, it has proven challenging to legislate whether online providers are supportive of students with different learning needs. For example, in a recent US Supreme Court case, the language that legally protects learners with disabilities was debated in part because lower courts interpreted the meaning of the law only to protect those learners to the extent of providing access to a learning experience that is better than "trivial" (Totenberg, 2017). Insufficient accessibility consideration for learners with disabilities is also a relevant problem in the Massive Open Online Courses (MOOCs) space. For example, the MOOC provider edX settled a lawsuit out of court with the US Department of Justice over the lack of support for learners with visual and hearing impairment (Duehren, 2015). In other words, although governments provide legislation and guidelines for accessible educational resources, there are concerns about the availability of resources to ensure inclusion and equality (Department for Education, 2017).

The Open University, GB

Corresponding author: Francisco Iniesto (francisco.iniesto@open.ac.uk)

The time spent designing educational resources that consider accessibility requirements should not require more general benefits to the broader population to justify the effort as all learners will benefit (de Waard et al., 2014). Indeed, the Universal Design for Learning (UDL) framework outlines that efforts toward improving accessibility have the potential to enhance the learning environment for all people (Hall, Meyer and Rose, 2012; Meyer and Rose, 2002). One place where we can and should improve the experience of learners is through the assistive technology of text-to-speech (TTS). TTS is a feature where the computer reads text aloud (Charlson, 2014) and is now widely available in most Office tools and smartphone applications, as well as in Internet browsers. For example, contemporary trends in the 2017 Consumer Electronics Show (CES) demonstrate that Amazon's Alexa was a dominant presence in emerging technology (Chris, 2017). In seeing an increasing number of devices adopting speech interfaces like Amazon's Alexa, we will likely see an increase in people interacting with computerised voices. By trying to understand the role TTS plays in the online learning context, we aim to contribute to a more comprehensive understanding of how people interact with computerised voices, and potentially achieve better learning outcomes for learners. In shifting from the perspective of providing better than "trivial" learning experience towards achieving the learning potential of learners with disabilities, we stand a chance of gaining insights into how this might relate to the broader population in education and beyond.

Although from a technological perspective an increasing number of tools and approaches have been developed

using TTS, there is limited research on how useful these TTS tools are in the emotional delivery of text through synthetic voices in the learning experience. In our initial explorative work (Hillaire, Iniesto and Rienties, 2017), we found that two TTS voices frequently expressed negative emotion, rarely expressed neutral emotion, and that the emotional expression was seldom aligned with the emotion detected in the text by human raters. This study builds on those findings by comparing two TTS voices with a synthetic voice that is configurable in terms of the potential to align emotional expression from synthetic voices with emotion identified in the text. First, we expand on those findings by providing detailed examples of the emotion identified in the text by human raters, and use those examples to discuss the alignment of TTS voices with the human detected emotion. Finally, we present new results of configuring a synthetic voice to express emotion aligned with the emotion identified in the text by human raters.

Designing Accessible MOOCs

The enrolment numbers of learners with disabilities are increasing for those who choose online education institutions for their studies (Slater et al., 2015). The lifelong learning paradigm integrates education, work and personal life in a continuous process and allows learners to access knowledge, competencies and skills, and develop them both personally and through work (Butcher and Rose-Adams, 2015). MOOCs are beneficial given their defining characteristics of openness within a structured learning framework, as well as their low costs of learning. Furthermore, MOOCs allow learners to plan their time at their preferred pace and place, allow opportunities for social learning, and provide a chance to gain new skills and knowledge (Scanlon, McAndrew and O'Shea, 2015).

As learning in many MOOCs is often self-directed, it requires a more significant commitment from a learner to develop and maintain an effective self-regulated learning strategy, with an in-depth research aptitude, reflexive capacity, and a high level of personal autonomy (Littlejohn et al., 2016; Sharples et al., 2012). MOOC design should consider each learner's abilities, learning goals, where learning takes place, and which specific devices the learner uses. Providing accessible MOOCs could offer the flexibility of learning and benefits to learners, irrespective of any disability. The importance of accessibility to online educational resources is widely acknowledged (Acosta and Luján-Mora, 2016), but often not considered when designing MOOCs (Iniesto et al., 2017a).

Text-to-Speech and Emotional Expression

One way to make MOOC learning materials accessible for learners with visual impairments is to use TTS. As part of the US National Center on Accessing the General Curriculum, research on TTS conducted by CAST.org (Strangman and Hall, 2003) examined 13 studies that used either TTS or tape recordings, highlighting that audio provided a benefit when supporting reading comprehension. In examining both human recordings and computerised voices, this report noted that they could only identify one small-scale study that took a quantitative method to compare the use of synthetic and human voices in learning; in this study (n = 3) in one case a synthetic voice produced the best learning outcomes, and in two cases human recorded voices produced the best learning outcomes. One potential contributing factor to this difference in performance for students between hearing a synthetic voice and hearing a human voice could be the difference in prosodic expression of emotion as it relates to how the student perceives the content.

Indeed, in a study with 60 people, Danion et al. (1995) provided evidence that valence of emotional words (i.e., positive, negative) had higher recall rates, whereas in recognition tasks more neutral words were recognised than negative words. In contrast, for word completion tasks, emotions did not appear to influence performance. Indeed, recent research indicated that TTS using synthetic verse recordings of humans reading text influenced the fundamental learning activity of reading comprehension (Citron et al. 2014).

A review on emotion detection technology found Vokaturi[1] to be a state-of-the-art detection of emotion from speech, with an accuracy rate of 66.5% (Garcia-Garcia, Penichet and Lozano, 2017). Vokaturi has been used for research on MOOCs to interpret voices from video lectures for highlighting lecture notes (Che, Yang and Meinel, 2018), indicating this technology is starting to be adopted by the MOOC context. More importantly, we are building on previous findings using Vokaturi (Hillaire et al. 2017) by expanding on work focused on emotional text and TTS.

In this study, the new synthetic voice we are introducing to the analysis is IBM Watson's[2] TTS functionality because it supports Speech Synthesis Markup Language (SSML) that can configure the voice to express text with specific emotional delivery (Pitrelli et al. 2006). With SSML, IBM Watson supports expressing text as good news and expressing text as an apology. We use these two settings to configure text identified as negative by human raters to be read as an apology and text identified as positive to be read as good news. We then investigate the potential to use SSML to improve the alignment between the emotional expression of a synthetic voice with the emotion identified in course text by human raters.

Current research indicates that educational content can be adapted for specific profiles of learners in online learning settings, and in particular in MOOCs (Sein-Echaluce, Fidalgo-Blanco and Garcia-Peñalvo, 2017). This can have benefits not only for how the information is shown to learners but also to provide recommendations for these learners (Iniesto, Rodrigo and Hillaire, 2019). Including how MOOC materials are perceived emotionally when using TTS seems a natural next step in the process of personalisation of the learning experience (Badia, Garcia and Meneses, 2019).

Study Context

Although both the use of TTS and the valence of words have demonstrated effects on word level tasks (such as acquisition and recall), to the best of our knowledge, there are no studies that have examined the intersection of

valence and TTS in the context of MOOCs. To start this work, we need to establish some form of an initial baseline. To do this for the emotional expression of course text in MOOCs, we first establish the emotional dimensions of text regarding word choice in the text using human raters, and then examine how that text is interpreted for expression through the accessibility feature of TTS. Both the word selection of what is communicated in the text as well as how those words are spoken become a subjective expression of the object of learning. As indicated in **Figure 1**, when the speaker of the text is neither the student nor the course designer, then the emotional interpretation of the speaker becomes a third factor in generating an emotional layer on the object of learning. In TTS software, the third factor is the implementation of the software interpreting text and the selection of expressive features that result in the prosody of speech.

When learning materials are delivered through TTS technology, we can identify the objects of learning delivered via the emotional package based on the valence of word selection in the text, as detected through human raters on the emotional content of the text, as well as the prosody of speech produced by TTS software. Therefore, to explore whether TTS "produces" valence (positive/negative) and emotional expression when reading written text out loud, we first had to determine the emotional expression of MOOC text using human raters (RQ1). Then we compared and contrasted the emotional expression of MOOC text with those expressed by TTS (RQ2). Finally, to improve the match between human raters and TTS, we explored how we could improve the configuration of the TTS voices (RQ3). In other words, the three research questions (RQs) for this study are:

- **RQ1:** To what extent does MOOC text contain emotional expression as indicated by two human raters?
- **RQ2:** To what extent does the emotional expression of synthetic voices in TTS align with emotion identified in the text by human raters?
- **RQ3:** To what extent can we configure synthetic voices to express emotion aligned with the emotional content in MOOC text?

Method

The sample included three of the leading MOOC providers considered in the list made by Shah (2017) following the criteria of representativeness and worldwide perspective, based on the use of English language and acknowledging an exploratory perspective (Twining, 2010). **Table 1** summarises the MOOCs selected to carry out our study, whereby we ensured that we selected three MOOCs from different specialisations (i.e. social sciences, physical science and engineering, and personal development) and from three distinct and commonly used MOOC platforms: FutureLearn, edX and Coursera.

The sample within MOOC web pages was based on Iniesto et al. (2017b) while sampling MOOCs for accessibility evaluations, as indicated in **Figure 2**:

- **Each course home page.** Including general information related to the course.
- **A discussion page.** These pages were used to allow discussion between learners, such as discussion forums.
- **An educational resource page.** These included activities such as watching a video or reading a text.

We used the screen reader NVDA[3] and CamStudio[4] to record the course interaction (both are open source software).

The three text samples from each MOOC were first preprocessed by the NLTK tokeniser[5] to break each sample into sentences. As the text from the course material was of variable length in terms of the number of sentences in each section (i.e., the home page, the discussion page, and the resource page), when converting the sample into sentences for analysis, we produced a different number of total sentences for each course. The 40 sentences for analysis comprised 13 for FutureLearn; 12 for edX; and 15 for Coursera. Each of the sentences had corresponding audio clips in two synthetic TTS voices. Voice-1 was the male voice whereas Voice-2 was the female one. Also, we generated audio clips using IBM Watson for the same sample to examine the efficacy of a synthetic voice that we configured for emotional expression aligned with emotion detected in the text by human raters.

Data Analysis

First, to address RQ1, the text of the 40 sentences was added to an Excel spreadsheet that contained the course, section, sentence number from the section, and sentence text. Two researchers (Author 1, Author 2) rated the positive content of each sentence on a scale from one (no emotion) to five (high positive emotion). The researchers

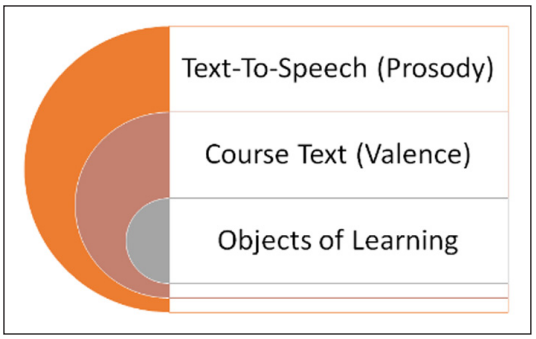

Figure 1: A measure of emotional learning material.

Table 1: Courses selected for the study.

Name of course	Institution	Provider
Caring for Older People: A Partnership Model	Deakin University	FutureLearn
Mechanics: Momentum and Energy	Massachusetts Institute of Technology	edX
Successful Negotiation: Essential Strategies and Skills	University of Michigan	Coursera

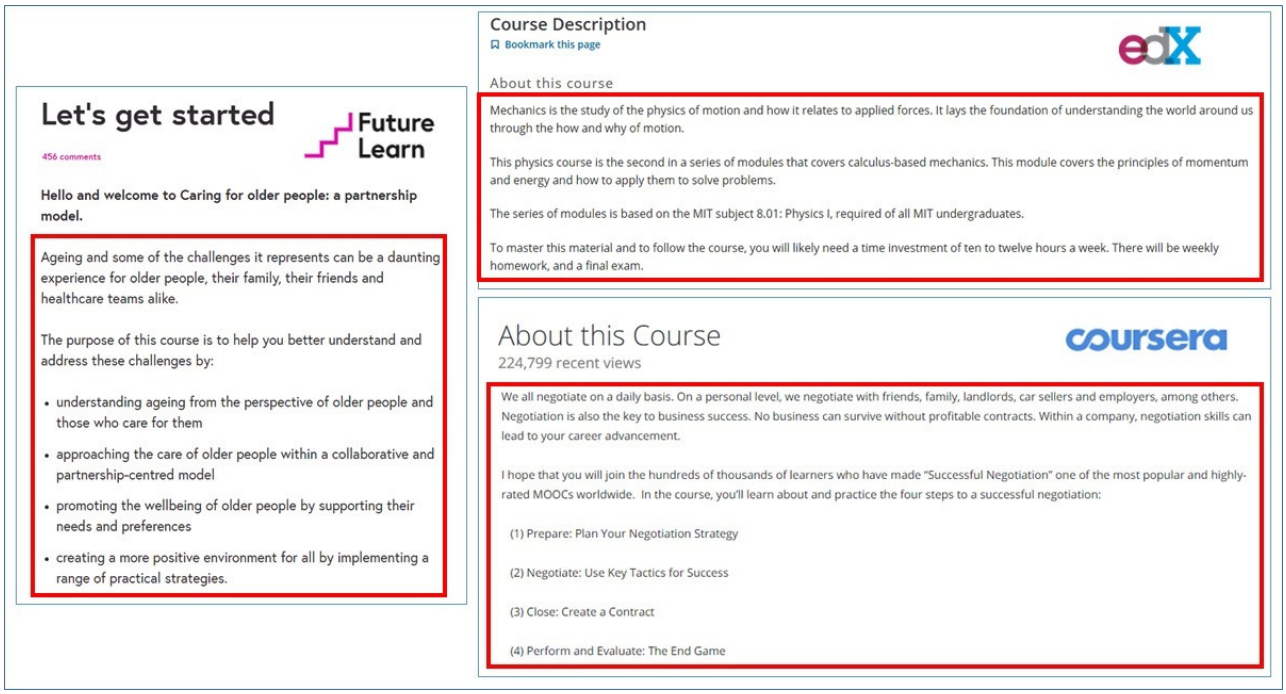

Figure 2: Exemplar home page of selected MOOCs.

coded a sentence between two and four if the sentence contained some positive sentiment. The same approach was used to indicate the amount of negative sentiment. The raters used a coding scheme developed for the social web that has an online six-hour training course.[6] Rater-1 went through the online course and explained the coding scheme to Rater-2. Second, the following ratings were classified as positive, negative, neutral or mixed using a peak method to determine the dominant sentiment. The rating was neutral if both positive and negative were rated as a one. The rating was positive if the positive score was greater than the negative score. The rating was negative if the negative score was greater than the positive score. The rating was mixed if the positive and negative score were both greater than one and equal to each other.

Third, to address RQ2 for each audio clip generated by a synthetic voice, the emotional expression of the audio recordings was detected using Vokaturi. Vokaturi is an open source application programming interface (API) that implemented the mining of prosodic features. Based on those features, this software compared the two databases of recordings of audio, where labels indicated specific emotions expressed. Vokaturi detects prosodic patterns similar to audio from the databases and predicts that the expression was of one of the emotions based on the labels from the databases. The emotions that Vokaturi can detect are "Neutral", "Happy", "Sad", "Angry" and "Fear". Vokaturi returned a prediction about the percentage likelihood of these five emotions. The predictions came back as values that added up to one, representing the probability of whether a particular emotion was present. We used a peak prediction score to determine the most likely emotional expression in the audio clip by taking the highest probability of each prediction to select a single emotion word that best described the valence of the audio recording. "Happy" predictions were considered positive valence. "Sad", "Angry" and "Fear" predictions were considered negative valence. "Neutral" predictions were considered neutral.

Finally, to address RQ3, we generated audio clips for the sample MOOC content analysed in using IBM Watson's TTS technology. IBM Watson's TTS supported the ability to configure the emotional expression of the synthetic voice using tags that indicated the intended emotional expression. For this study, we generated audio for our sample using the "express-as" feature of IBM Watson that supported generating audio expressed as an apology or expressed as good news. The intent of using these two tags was to find out if we could generate audio with intended emotional expression using TTS technology. We investigated if configuring TTS to express good news resulted in audio that is positive. Additionally, we checked if expressing as an apology would result in audio that is negative. We again used Vokaturi to analyse the audio to generate comparable results with our analysis of TTS. We calculated precision, recall and f-measure for each valence category and macro averaging using Scikit-learn.

Results

RQ1: To what extent does MOOC text contain emotional expression as indicated by two human raters?

We found a consensus in the sample of MOOC text for 31 out of 40 sentences (77.5%), and as there were many zero cells in the confusion matrix (see **Table 2**), which causes problems with computing Kappa statistics for agreement (Yarnold, 2016), we used Bennett, Alpert and Goldstein's (1954) S, which was 0.70 indicating substantial agreement (Landis and Koch, 1977). We did not resolve disagreement

Table 2: Confusion matrix for human raters.

		Rater-2			
		Positive	**Negative**	**Neutral**	**Mixed**
Rater-1	Positive	7	0	0	0
	Negative	0	7	0	0
	Neutral	5	4	17	0
	Mixed	0	0	0	0

ratings and rather evaluated Rater-1, Rater-2 and consensus ratings when exploring RQ2 because ultimately the end user of course material is an individual. The three evaluations provide insight into what might depict a general experience (using consensus ratings) and what might depict an individual experience (using Rater-1 and Rater-2). According to **Table 2**, the most frequently classified category was a neutral expression, and none of the ratings of the sentences produced a single outcome rated as mixed emotion.

Rater-1 identified 26 messages as neutral whereas Rater-2 identified 17. The nine items of disagreement appeared to indicate that Rater-1 found more sentences to be neutral as indicated in **Table 2**. Both raters found 35% of text samples (14 out of 40) to contain emotion.

In terms of RQ1, our results suggested that somewhere between 35% and 57% of MOOC sentences contained an emotional expression. By considering the agreement of two raters, 35% of MOOC text from our sample of three MOOCs contained an emotional expression. Our results suggest that about one out of every three sentences (35%) in a MOOC contained emotional text.

RQ2: To what extent does the emotional expression of synthetic voices in TTS align with emotion identified in the text by human raters?

We recorded audio clips generated by two screen readers (one male voice and one female voice) using all 40 MOOC sentences from our sample. The audio clips were analysed using Vokaturi, which was able to predict the emotional expression for 39 MOOC sentences. One sentence, when read by the screen reader, resulted in an audio clip with insufficient data for Vokaturi to make a prediction.

We compared the agreement of emotional expression for Voice-1 (male) and Voice-2 (female), which resulted in the agreement of emotional expression 27 out of 39 times (69% agreement). As there were many zero cells in the categories that cause problems with computing Kappa statistics (Yarnold, 2016) for agreement, we used Bennett et al.'s (1954) S, which was 0.59 indicating moderate agreement (Landis and Koch, 1977). Voice-2 was almost always predicted to have a negative emotion, with 38 out of 39 recordings considered as negative (the remaining one audio file considered neutral). Voice-1 was predicted to be positive about 33% of the time (12 out of 39 times) and negative 66% of the time (27 out of 39 times) as indicated in **Table 3**. In other words, while the human raters found a high frequency of neutral content (RQ1), the audio detection of emotion features most frequently identified negative expression.

Using the two human raters, we calculated the precision, recall and f-measures considering the ratings and the consensus of the raters as true scores. The emotional expression categories were calculated separately, and the weighted average was computed across negative, neutral and positive using macro averaging to give a summary computation that evaluated the agreement. Given that the two voices exhibited differentiated outcomes as measured by the emotion prediction from the audio clips (see **Table 3**), the same calculations were made for each voice separately. These evaluations are summarised in **Tables 4** and **5** in which the results are given as precision (P), recall (R) and f-measure (F).

For neutral expression Voice-1 as compared with Rater-2 and consensus had precision, recall and f-measure scores of 0.00, indicating it is never expressing neutral when either Rater-2 or the consensus of Rater-2 and Rater-1 considered the text neutral. In other words, there appeared to be a very low level of agreement between Rater-1 and Voice-1 on a neutral expression, as there was a precision of 1.00, recall of 0.04, and f-measure of 0.07. These statistics suggested that Voice-1, which was detected to be neutral once (see **Table 3**), was in agreement with Rater-2 that time. However, Voice-1 was more frequently not detected to be neutral when Rater-1 considered the text to be neutral, which occurs 26 times (see **Table 2**), resulting in the low recall of 0.04 (1/26) and f-measure of 0.07.

Voice-1 is never detected to be positive (see **Table 3**), which explains why the precision, recall and f-measures were all 0.00 when comparing the emotion detected in Voice-1 as compared with human raters identifying text as positive (Rater-1: 7 times; Rater-2: 12 times; and consensus: 7 times – see **Table 2**). Voice-2 performed at or above chance levels (0.33) for recall when comparing with human raters. The highest accuracy occurred when comparing with consensus rating with an f-measure of 0.50 as compared with the f-measure of 0.42 for both Rater-1 and Rater-2.

When considering text identified as negative, it is important to note that Voice-1 had low precision across Rater-1 (0.18), Rater-2 (0.26) and consensus (0.23). Similarly, Voice-2 had low precision across Rater-1 (0.19), Rater-2 (0.26) and consensus (0.24). In contrast, the recall was very high for Voice-1 compared with Rater-1 (1.00), Rater-2 (0.91) and consensus (1.00). Similarly, Voice-2 had a high recall compared with Rater-1 (0.71), Rater-2 (0.64)

Table 3: Confusion matrix for synthetic voices.

		Voice-1			
		Positive	**Negative**	**Neutral**	**Mixed**
Voice-2	Positive	0	11	1	0
	Negative	0	27	0	0
	Neutral	0	0	0	0
	Mixed	0	0	0	0

Table 4: Evaluation of Voice-2 emotional expression compared to human ratings.

	Negative			Neutral			Positive			All (macro average)		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Rater-1 n = 39	0.19	0.71	0.29	0.00	0.00	0.00	0.33	0.57	0.42	0.17	0.43	0.24
Rater-2 n = 39	0.26	0.64	0.37	0.00	0.00	0.00	0.42	0.42	0.42	0.23	0.35	0.26
Consensus n = 31	0.24	0.71	0.36	0.00	0.00	0.00	0.44	0.57	0.50	0.23	0.43	0.29

Table 5: Evaluation of Voice-1 emotional expression compared to human ratings.

	Negative			Neutral			Positive			All (macro average)		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Rater-1 n = 39	0.18	1.00	0.30	1.00	0.04	0.07	0.00	0.00	0.00	0.39	0.35	0.13
Rater-2 n = 39	0.26	0.91	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.30	0.13
Consensus n = 31	0.23	1.00	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.33	0.12

and consensus (0.71). Considering the consistent patterns of low precision with high recall, we interpret this to mean that both Voice-1 and Voice-2 were likely overly negative, causing a high level of recall with a low level of precision. This was consistent with the frequency of Voice-1 detected as negative 38 out of 39 times and Voice-2 27 out of 39 times (see **Table 3**).

The low precision and high recall also described the overall statistics reported in the all (macro average) column of **Tables 4** and **5**. When considering Voice-2 as compared with consensus ratings, there was a precision of 0.23 and recall of 0.43 (above chance levels). The same was true for Voice-1 compared with consensus ratings, with a precision of 0.08 and recall of 0.33 (at chance levels). The key takeaway from these results is that the synthetic voices are both likely overly negative when considering their expression as compared to the emotional content of the text they are reading.

RQ3: To what extent can we configure synthetic voices to express emotion aligned with the emotional content in MOOC text?

For the 31 text samples from our selected MOOCs where we had a consensus for emotional rating, we used IBM Watson's TTS to generate audio clips configured to express emotion in the audio clip aligned with the emotion identified in the text. We used the "express-as" feature of IBM Watson to express positive text as good news; negative text as an apology; and neutral text with no expression setting. This improved the results reported in RQ2 as illustrated in **Table 6**.

IBM Watson was configured to express the emotion aligned with the consensus rating on the emotion in the text. The findings indicated in **Table 6** found that IBM Watson performed best when compared with Voice-1 and Voice-2 from TTS, as indicated with an f-measure of 0.31 for IBM Watson as compared to Voice-2's f-measure of 0.29 and Voice-1's f-measure of 0.12.

Given that we had three categories of positive, negative and neutral, the statistic chance levels of accuracy (randomly guessing) would have a recall of 0.33. In terms of recall of Voice-1, the recall of 0.33 indicated performance at chance levels. Both Voice-2 with a recall of 0.43 and IBM Watson with a recall of 0.55 performed above chance levels. In terms of precision, IBM Watson's score of 0.45 outperformed Voice-1 (0.08) and Voice-2 (0.23).

We can see by the f-measure scores calculated for each valence category (negative, neutral and positive) that IBM Watson did better than Voice-1 and Voice-2 at negative and neutral expression. Furthermore, IBM Watson did better than Voice-1 at positive expression, but Voice-2 performed best at positive expression. However, IBM Watson was considered to be the best overall at matching expression with the emotional content of text for all three categories of negative, neutral and positive when compared with Voice-1, and it performed better than Voice-2 at negative, and neutral expression. This makes sense given that we configured IBM Watson to express emotion based on human labels. Although IBM Watson performed better, it is still far from perfect, demonstrated by its overall recall of 0.55. Also, all of the voices appeared to do poorly at matching the emotional content of the text when the text was considered neutral. The key takeaway from these results is that synthetic voices in their current state performed poorly for neutral communication.

Discussion and Future Research

An increasing number of learners are using TTS technologies to learn from written and online materials, including the materials from MOOCs. Our research explored

Table 6: Evaluation of IBM Watson "express-as" feature to generate audio with emotion.

	Negative			Neutral			Positive			All		
	P	R	F	P	R	F	P	R	F	P	R	F
Voice-1	0.23	1.00	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.33	0.12
Voice-2	0.24	0.71	0.36	0.00	0.00	0.00	0.44	0.57	0.50	0.23	0.43	0.29
IBM Watson	0.43	0.38	0.40	0.06	1.0	0.11	0.86	0.27	0.41	0.45	0.55	0.31

whether the use of such TTS technologies had a positive or negative impact on (un)expressed emotions in speech and text comparing a sample of 40 sentences from three commonly available MOOCs. When investigating the extent to which MOOC texts contained emotional expression, we found about one in three sentences (35%) had emotional expression. This pilot established that there was an emotional layer that was added by TTS technologies to the content of MOOC course material when text expression was used. As the emphasis of analysis for this study was on the measurement of emotion, these results established the (perhaps unexpected) presence of emotion in MOOCs. Of course, our study did not provide insights into the influences emotional expression may (or may not) have on learning when users use such TTS technologies. However, as highlighted in our literature review, multiple studies demonstrated that emotional expression influenced tasks, such as recognition and recall of words (Citron et al., 2014; Danion et al., 1995). Considering that 35% of the selected texts in our sample were emotional expression, future research should consider what influence emotional expression has on student performance in MOOCs.

When investigating if TTS voices had emotional expression that aligned with the emotion of MOOC text (as identified by two human raters), we found that there was a relatively poor alignment due to TTS being overly negative when reading the selected text. This disconnect may be an explanatory factor as to why the experience of listening to humans reading a text for some learning tasks is superior to hearing synthetic voices reading the same text (Strangman and Hall, 2003). Future research should also consider how the (mis)alignment of TTS emotional expression influences reading comprehension, in order to evaluate the extent to which this factor has the potential to influence learning. As highlighted by RQ3, with the attempts to configure a synthetic voice to align its emotional expression with the emotion detected in the text, we saw an improvement. However, the extent to which that improvement has the potential to influence learning is highly dependent on the extent to which (mis)alignment of TTS to course text influences learning.

As synthetic voices become more ubiquitous, educational researchers can and possibly should investigate the impact of emotional expression of TTS on learning. As technologies like Amazon's Alexa expand their reach into our everyday lives (Chris, 2017), they may either be directly adopted by schools or merely be used by students. The emergence of using synthetic voices like those in assistive technology is an example of how improving accessibility can have the potential to benefit everyone (Hall et al. 2012; Meyer and Rose, 2002). However, if educational researchers do not investigate the implications of emotional expression in the course text and the (mis)alignment of synthetic voice expression for learning, then adoptions of these technologies will introduce an unknown effect on student learning. Although the results of this pilot study are limited by a small sample of course material, only two human raters, and only two synthetic voices, these results slightly reframe the problem by highlighting that current TTS technologies appear to express emotion. Future research should examine the effect that emotional expression in TTS has on learning.

One of the limitations in the current research includes, as shown while answering RQ3, work remains to create SSML technologies that deliver the intended emotional expression before we can perfectly align synthetic voices with emotion in text. There is also a limitation to the use of MOOCs in English. Further research needs to be done in an emerging global educational technology to provide multilingual support. Furthermore, we acknowledge that the sample is small, it cannot be considered representative of MOOCs, but—in the context of this pilot study—our results suggest that MOOCs might have an emotional expression in a comparable rate. As a pilot study, we found that there is enough emotional expression with the small sample size to warrant a more systematic investigation to identify the scope of emotional expression in MOOCs.

The technical barriers to delivering emotional content have been lowered with the adoption of Speech Synthesis Markup Language (SSML) processing in TTS technologies. SSML is a markup language that allows authors to indicate the prosodic expression of the text: some TTS technologies are implemented to interpret SSML to vary the prosodic expression giving the authors of text more control over how the TTS reads the text emotionally.

While there is a potential impact that synthetic voices may have on all students in the future, there are students today that depend on assistive technologies. At The Open University (OU), nearly 16% of the population are self-identified as students with disabilities showing a gap in the pass rate for disability declared (Rienties et al. 2016). In this population, more than a thousand students self-reported a disability related to sight. As educational researchers, we can focus on emerging trends in technology, learn from the overlap with populations who are early users of that technology due to their need for accessibility and so gain insights into how best to support students.

Future research will include further exploration of the audio predictions for the validity of students' interviews after listening to OU context recordings. The objective is to understand if the variable emotional expression in online courses' text and the variable emotional expression of synthetic voices are influencing learning outcomes. As human raters found emotional expression in course material text, future research should explore how emotional expression in course material influences student participation. For example, does positive and encouraging language increase student participation? The relationship between emotional expression in course material and the potential influence on student participation should examine effects for students reading the text and those accessing course material through a screen reader.

Notes

[1] Vokaturi, https://vokaturi.com/.
[2] IBM-Watson, https://www.ibm.com/watson.
[3] NVDA, https://www.nvaccess.org.
[4] CamStudio, https://camstudio.org/.
[5] NLTK tokeniser, https://www.nltk.org/.
[6] Scikit-learn, https://scikit-learn.org/stable/index.html.

Acknowledgements

This work is supported by two Leverhulme Trust Doctoral Scholarships in Open World Learning based in the Institute of Educational Technology (IET) at The Open University. Garron thanks the Lummi Nation for supporting his work through the Lummi Higher Education Grant; Francisco thanks OpenTEL and the Global OER Graduate Network (GO-GN), which is supported by the William and Flora Hewlett Foundation.

Competing Interests

The authors have no competing interests to declare.

References

Acosta, T and **Luján-Mora, S.** 2016. Comparison from the levels of accessibility on LMS platforms that supports the online learning system. In: *8th International Conference on Education and New Learning Technologies*, Barcelona, Spain, 4–6 July. DOI: https://doi.org/10.21125/edulearn.2016.1579

Badia, A, Garcia, C and **Meneses, J.** 2019. Emotions in response to teaching online: Exploring the factors influencing teachers in a fully online university. *Innovations in Education and Teaching International*, 1–12.

Bennett, EM, Alpert, R and **Goldstein, AC.** 1954. Communications through limited-response questioning. *Public Opinion Quarterly*, 18(3): 303–308. DOI: https://doi.org/10.1086/266520

Butcher, J and **Rose-Adams, J.** 2015. Part-time learners in open and distance learning: revisiting the critical importance of choice, flexibility and employability. *Open Learning: The Journal of Open, Distance and e-Learning*, 30(2): 127–137. DOI: https://doi.org/10.1080/02680513.2015.1055719

Charlson, B. 2014. Accessible Technology Options for the Blind and Visually Impaired Reader. Available at https://www.youtube.com/watch?v=BzpEtheQwhs [Accessed 22 July 2019].

Che, X, Yang, H and **Meinel, C.** 2018. Automatic Online Lecture Highlighting Based on Multimedia Analysis. *IEEE Transactions on Learning Technologies*, 11(1): 27–40. DOI: https://doi.org/10.1109/TLT.2017.2716372

Chris, B. 2017. CES 2017: Amazon's virtual aide Alexa shouts above rivals. *BBC*, 7 January. Available at https://www.bbc.co.uk/news/technology-38539326 [Accessed 22 July 2019].

Citron, FMM, Gray, MA, Critchley, HD, Weekes, BS and **Ferstl, EC.** 2014. Emotional valence and arousal affect reading in an interactive way: Neuroimaging evidence for an approach-withdrawal framework. *Neuropsychologia*, 56(1): 79–89. DOI: https://doi.org/10.1016/j.neuropsychologia.2014.01.002

Danion, JM, Kauffmann-Muller, F, Grang, D, Zimmermann, MA and **Greth, P.** 1995. Affective valence of words, explicit and implicit memory in clinical depression. *Journal of Affective Disorders*, 34(3): 227–234. DOI: https://doi.org/10.1016/0165-0327(95)00021-E

Department for Education. 2017. Inclusive teaching and learning in higher education. Available at https://www.gov.uk/government/publications/inclusive-teaching-and-learning-in-higher-education [Accessed 22 July 2019].

de Waard, I, Gallagher, MS, Zelezny-Green, R, Czerniewicz, L, Downes, S, Kukulska-Hulme, A and **Willems, J.** 2014. Challenges for conceptualising EU MOOC for vulnerable learner groups. In: *Proceedings of the European MOOC Stakeholder Summit 2014*, University of Graz, Austria, 22–24 February, pp. 33–42.

Duehren, AM. 2015. EdX Settles With Department of Justice. *The Harvard Crimson*, 3 April. Available at http://www.thecrimson.com/article/2015/4/3/edX-settles-department-justice/ [Accessed 22 July 2019].

Garcia-Garcia, JM, Penichet, VMR and **Lozano, MD.** 2017. Emotion detection: a technology review. In: *Proceedings of the XVIII International Conference on Human Computer Interaction*, Cancun, Mexico, 25–27 September, Article No. 8. DOI: https://doi.org/10.1145/3123818.3123852

Hall, TE, Meyer, A and **Rose, DH.** (eds.) 2012. *Universal Design for Learning in the Classroom*. Guilford Press.

Hillaire, G, Iniesto, F and **Rienties, B.** 2017. Toward Emotionally Accessible Massive Open Online Courses (MOOCs). In: *14th AAATE Congress 2017*, Sheffield, 13–14 September.

Iniesto, F, McAndrew, P, Minocha, S and **Coughlan, T.** 2017a. An investigation into the perspectives of providers and learners on MOOC accessibility. In: *TEEM 2017 Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality*, Cádiz, Spain, 18–20 October, Article No. 95. DOI: https://doi.org/10.1145/3144826.3145442

Iniesto, F, McAndrew, P, Minocha, S and **Coughlan, T.** 2017b. Auditing the accessibility of Massive Open

Online Courses (MOOCs). In: *14th AAATE Congress 2017*, Sheffield, 13–14 September.

Iniesto, F, Rodrigo, C and **Hillaire, G.** 2019. Applying UDL principles in an inclusive design project based on MOOCs reviews. In: Gronseth, SL and Dalton, EM (eds.), *Universal Access Through Inclusive Instructional Design: International Perspectives on UDL*. New York: Routledge (in press).

Landis, JR and **Koch, GG.** 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1): 159–174. DOI: https://doi.org/10.2307/2529310

Littlejohn, A, Hood, N, Milligan, C and **Mustain, P.** 2016. Learning in MOOCs: Motivations and self-regulated learning in MOOCs. *The Internet and Higher Education*, 29: 40–48. DOI: https://doi.org/10.1016/j.iheduc.2015.12.003

Meyer, A and **Rose, DH.** 2002. *Teaching Every Student in the Digital Age: Universal Design for Learning*. Alexandria, VA: Association for Supervision and Curriculum Development.

Miles, S. 2000. Overcoming Resource Barriers: The Challenge of Implementing Inclusive Education in Rural Areas. *Enabling Education Network*. Available at https://www.eenet.org.uk/resources/docs/bonn_1.docx [Accessed 22 July 2019].

Pitrelli, JF, Bakis, R, Eide, EM, Fernandez, R, Hamza, W and **Picheny, MA.** 2006. The IBM Expressive Text-to-Speech Synthesis System for American English. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4): 1099–1108. DOI: https://doi.org/10.1109/TASL.2006.876123

Rienties, B, Edwards, C, Gaved, M, Marsh, V, Herodotou, C, Clow, D, Cross, S, Coughlan, T, Jones, J and **Ullmann, T.** 2016. Scholarly insight 2016: a Data wrangler perspective. Available at http://oro.open.ac.uk/48244/.

Scanlon, E, McAndrew, P and **O'Shea, T.** 2015. Designing for educational technology to enhance the experience of learners in distance education: how open educational resources, learning design and MOOCs are influencing learning. *Journal of interactive Media in education*, *2015*(1). DOI: https://doi.org/10.5334/jime.al

Sein-Echaluce, ML, Fidalgo-Blanco, Á and **García-Peñalvo, FJ.** 2017. Adaptive and cooperative model of knowledge management in MOOCs. In: Zaphiris, P and Ioannou, A (eds.), *Learning and Collaboration Technologies. Technology in Education. 4th International Conference, LCT 2017*. Held as Part of HCI International 2017, Vancouver, BC, Canada, 9–14 July, Proceedings, Part I, pp. 273–284. Switzerland: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-58509-3_22

Shah, D. 2017. A product at every price: A review of MOOC stats and trends in 2017. *Class Central*. Available at https://www.classcentral.com/report/moocs-stats-and-trends-2017/ [Accessed 22 July 2019].

Sharples, M, McAndrew, P, Weller, M, Ferguson, R, Fitzgerald, E, Hirst, T, Mor, Y, Gaved, M and **Whitelock, D.** 2012. *Innovating Pedagogy 2012: Open University Innovation Report 1*. Milton Keynes: The Open University.

Slater, R, Pearson, VK, Warren, JP and **Forbes, T.** 2015. Institutional change for improving accessibility in the design and delivery of distance learning – the role of faculty accessibility specialists at The Open University. *Open Learning: The Journal of Open, Distance and e-Learning*, 30(1): 6–20. DOI: https://doi.org/10.1080/02680513.2015.1013528

Strangman, N and **Hall, T.** 2003. *Text Transformations*. Wakefield, MA: National Center on Accessing the General Curriculum.

Totenberg, N. 2017. Supreme Court Considers How Schools Support Students with Disabilities. *National Public Radio (NPR)*, 11 January. Available at https://www.npr.org/2017/01/11/509179589/supreme-court-considers-how-schools-support-students-with-disabilities?t=1560526776157 [Accessed 22 July 2019].

Twining, P. 2010. Educational information technology research methodology: looking back and moving forward. In: McDougall, A, Murnane, J, Jones, A and Reynolds, N (eds.), *Researching IT in education: Theory, practice and future directions*. London; New York: Routledge, pp. 153–168.

Yarnold, PR. 2016. ODA vs. π and κ: Paradoxes of kappa. *Optimal Data Analysis*, 5: 160–161.

How to cite this article: Hillaire, G, Iniesto, F and Rienties, B. 2019. Humanising Text-to-Speech Through Emotional Expression in Online Courses. *Journal of Interactive Media in Education,* 2019(1): 12, pp. 1–9. DOI: https://doi.org/10.5334/jime.519

Submitted: 15 February 2019 **Accepted:** 28 June 2019 **Published:** 10 September 2019

]u[*Journal of Interactive Media in Education* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS