Toward Emotionally Accessible Massive Open Online Courses (MOOCs)

Garron Hillaire[1], Francisco Iniesto and Bart Rienties
The Open University
{garron.hillaire, francisco.iniesto, bart.rienties} @open.ac.uk

Abstract. This paper outlines an approach to evaluating the emotional content of three Massive Open Online Courses (MOOCs) using the affective computing approach of prosody detection on two different text-to-speech voices in conjunction with human raters judging the emotional content of course text. The intent of this work is to establish the potential variation on the emotional delivery of MOOC material through synthetic voice.

Keywords. Emotions, Accessibility, MOOC

1. Introduction

In the field of education there are various considerations for disabled learners that examine how to make learning material accessible. However, one of the most frequently mentioned barriers that prevents making materials accessible is the perceived lack of resources even in highly resourced environments [1]. In a recent US Supreme Court case the language that legally protects disabled learners was debated in part because lower courts have interpreted the meaning of the law to only protect those learners so far as to provide access to a learning experience that is better than "trivial" [2]. Insufficient considerations for disabled learners is a relevant problem in the MOOC space as illustrated by edX settling a law suit with the US Department of Justice out of court over the lack of support for students with visual and hearing impairment [3]. One place where we can and should learn from the fringe population is through the assistive technology of text-to-speech (TTS). TTS is a feature were the computer reads text aloud [4]. By understanding the role TTS plays in the learning at scale context we can contribute to better learning outcomes for learners and contribute to a more comprehensive understanding of how people interact with computerized voices.

The intent of this work is to establish the potential variation on the emotional delivery of MOOC material. This variation will be based on the baseline of emotional expression in course text as detected by human raters. As this work is in early stages the current analysis is focused on the human rating of text as it compares to two potential voices in text-to-speech: a male voice, and a female voice. The results indicate that different TTS voices produce audio files that are predicted to have different emotional content when reading the same text. Moreover, the predicted emotional content does not align well with how two human raters interpreted the same text.

2. Methods

The courses selected to carry out the study are all from different disciplines (social sciences, physical science and engineering, and personal development) and each is presented in one of the following MOOC platforms: FutureLearn (Caring for Older People: A Partnership Model, Deakin University), edX (Mechanics: Momentum and Energy, MIT), and Coursera (Successful Negotiation: Essential Strategies and Skills, University of Michigan). A selection across disciplines and platforms was used to increase the variability of the sample. A selection of a set of Web pages was made as a representative example of each course: the homepage, a discussion page, and an activity page. A total of two different computerized voices (male and female) were used per website resulting in a total of 18 recordings. The 3 text samples from each course were pre-processed by NLTK tokenizer to break each sample into sentences. This took the initial 9 text samples from the courses and converted them into 40 sentences for analysis (13 sentences in Caring for Older People: A Partnership Model; 12 sentences in Mechanics: Momentum and Energy; 15 sentences in Successful Negotiation: Essential Strategies and Skills). The audio clips were then split into files that corresponded to the 40 sentences resulting in 40 text samples and 80 audio recordings. Voice-1 was designated as a male voice while Voice-2 was designated as a female voice.

To examine the emotional expression of course text in MOOCs first the valence of word selection in text was rated by researchers and second the valence of audio recordings generated from the two voices was predicted based on prosodic features of speech.

When coding text, human raters used an established set of coding instructions [6] that were modified from the twitter context by removing guidance for emoticons and abbreviations. The content in the sentence is rated on a scale from 1 to 5 on the presence of positive sentiment. Allocating 1 if the sentence contains no positive sentiment, and allocating 5 if the sentence contains very strong positive sentiment. Allocating a number between 2 and 4 if the sentence contained some positive sentiment. A similar scale was used to indicate the amount of negative sentiment. Then the ratings were classified as positive, negative, neutral, or mixed using a peak method to determine the dominant sentiment. The rating was considered neutral if both positive and negative were rated as a 1. The rating was considered positive if the positive score was greater than the negative score. The rating was considered negative if the negative score was greater than the positive score. The rating was considered mixed if the positive and negative score were both greater than 1 and equal to each other. These coding instructions were used by novice raters.

For prosodic detection the technology Vokaturi was used which extracts the primary dimensions of pitch, intensity, and spectral slope when detecting emotion [5]. Those features are compared to two annotated databased of audio recordings in order to predict if an audio recording expresses happiness, sadness, anger, fear, and neutrality. The predictions come back as values that add up to 1 representing the probability that the emotion is present. By taking the highest probability of each prediction a single emotion word is used to determine the peak valence of the audio recording. Happy predictions were considered positive valence. Sad, Angry, and Fear predictions were

considered negative valence. Neutral predictions remained neutral. The detection of emotion using this technology has been cross validated as %66.5 accurate.

3. Results

When the two raters examined the 40 sentences none of the ratings of these sentences produced a single outcome rating of mixed. The rating categories were compared for interrater reliability. There was consensus for 31 out of 40 sentences (%77.5 of the time) and the Cohens Kappa was 0.63891 which is considered substantial agreement. The 9 items of disagreement appear to indicate that Rater-1 found more sentences to be neutral. The most frequently classified category was neutral expression. Rater-1 identified 26 messages as neutral while Rater-2 identified 17 messages as neutral. This was nearly 2/3 of the ratings from Rater-1 and nearly half of the ratings from Rater-2. When examining the valence categories of the voice recordings the dominant expression was negative. Voice-1 was classified as negative 27 out of 39 times (roughly %69 of the time). Voice-2 was classified as negative an astounding 38 out of 39 times (roughly %97 of the time). While Vokaturi may have some limitations in accuracy, these results could indicate that the two synthetic voices selected for this study have a tendency to sound negative. The valence ratings for audio recordings were compared using interrater reliability. Agreement was detected from the voices 27 out of 39 times (%69.23 of the time). The Cohens Kappa was .05454 which is considered slight agreement.

While the human Raters found a high frequency of neutral content, the audio detection of emotion features most frequently identified negative expression. By comparing human raters with the prediction of the emotional content detected from the audio clip the goal is to identify how many of the emotion predictions about the audio align with the emotional content of the text. Given two raters the precision, recall, and f-measures were calculated using rater-1, rater-2, and the consensus of the two raters as true scores for text sentiment. The emotion expression categories of positive, negative, and neutral were calculated separately and the weighted average was computed across all three to give a summary computation that evaluates the agreement.

Table 1 – Evaluation of Voice-1 and 2 (Results given as precision P, recall R and f-measure F)

Voice 1	Positive			Negative			Neutral			All		
	P	R	F	P	R	F	P	R	F	P	R	F
Rater-1	0.33	0.57	0.42	0.19	0.71	0.29	0.00	0.00	0.00	0.09	0.23	0.13
Rater-2	0.42	0.42	0.42	0.26	0.64	0.37	0.00	0.00	0.00	0.20	0.31	0.23
Consensus	0.44	0.57	0.50	0.24	0.71	0.36	0.00	0.00	0.00	0.16	0.30	0.20
Voice 2	Positive			Negative			Neutral			All		
	P	R	F	P	R	F	P	R	F	P	R	F
Rater-1	0.00	0.00	0.00	0.18	1.00	0.30	1.00	0.04	0.07	0.68	0.20	0.10
Rater-2	0.00	0.00	0.00	0.26	0.91	0.40	0.00	0.00	0.00	0.07	0.25	0.11
Consensus	0.00	0.00	0.00	0.23	1.00	0.37	0.00	0.00	0.00	0.05	0.23	0.08

When evaluating the results of table 1 it is interesting to note the lack of accuracy in the identification of neutral expression. Table-1 illustrates in the neutral category that the one audio clip identified as neutral was in agreement with the human Rater-1 and yet on

the human rating side this was not one of the consensus ratings. This result may indicate that for these two synthetic voices it is difficult to express content in a neutral manner. Voice-2 was the only voice to achieve an expression that was predicted to be neutral, but it may have come at a cost. Voice-2 did not express any of the text as positive. Voice-2 did have high recall on negative expression, but that is likely explained by the fact that nearly all of the audio from Voice-2 was negative. Voice-1 did not achieve the same level of recall of negative expression, but it had a better balance of recall between positive and negative expression. Overall Voice-2 was predicted to be negative while Voice-1 was a blend of positive and negative. Neither voice did a particularly good job at aligning with the predictions of the emotional content of the text.

4. Discussion

The main driving force behind this investigation is to determine if the emotional expression of a synthetic voice influences the performance of learners. This paper investigates measuring the emotional expression in 3 MOOC courses when samples of the course text are read through text-to-speech. While there is clear discrepancy outlined in the audio and textual predictions of emotional content, the role of measurement error has not been exhaustively explored. In the next phase of this work the audio predictions will be further explored for validity in this context with user interviews after listening to recordings. There reason to further investigate both the variable emotional expression in course text and the variable emotional expression of synthetic voices as they have the potential to influence learning outcomes. While this illustrates a dimension that could be either a barrier or an asset, the next step needs to involve users of this technology to understand where this investigation should be prioritized.

Acknowledgements

This work is supported by two Leverhulme Trust Doctoral Scholarships in Open World Learning based in the Centre for Research in Education and Educational Technology at The Open University. Garron would like to thank the Lummi Nation for supporting his work; Francisco would like to thank the Global OER Graduate Network (GO-GN) which is supported by the William and Flora Hewlett Foundation.

References

1. Milesx, S. (2014). Overcoming Resources Barriers: The Challenges of Implementing Inclusive Education in Rural Areas. *Enabling Education Network*.
2. Totenberg, N. (2017). Supreme Court Considers How School suport Students with Disabilities *NPR*.
3. Duehren, A. M. (2015). EdX Settles With Department of Justice. *The Harvard Crimson*, 2015. A. M. Retrieved February, 2017, from: http://www.thecrimson.com/article/2015/4/3/edX-settles-department-justice/.
4. Charlson, B. (2014). Accessible Technology Options for the Blind and Visually Impaired Reader
5. Measuring acoustic features, Retrieved May, 2017, from: http://bit.ly/2rPrxqd.
6. Twitter Emotion Coding Instructions (2013), retrieved December 2016, from: http://sentistrength.wlv.ac.uk/documentation/TwitterVersionOfSentimentCodeBook.pdf