Dec 11th, 12:00 AM

AI-Generated Voice in Short Videos: A Digital Consumer Engagement Perspective

Jihao Luo
National University of Singapore, E0576164@u.nus.edu

Chenxu Zheng
National University of Singapore, chenxu@comp.nus.edu.sg

Jiamin Yin
Renmin University of China, yinjiamin@rmbs.ruc.edu.cn

Hock-Hai Teo
National University of Singapore, disteohh@nus.edu.sg

AI-Generated Voice in Short Videos: A Digital Consumer Engagement Perspective

Short Paper

Jihao Luo
Department of Information Systems and Analytics
National University of Singapore
15 Computing Drive, Singapore, 117418
jihaoluo@comp.nus.edu.sg

Chenxu Zheng
Department of Information Systems and Analytics
National University of Singapore
15 Computing Drive, Singapore, 117418
chenxu@comp.nus.edu.sg

Jiamin Yin
Department of Management Science and Engineering, School of Business
Renmin University of China
59 Zhongguancun Road, Beijing, China, 100872
yinjiamin@rmbs.ruc.edu.cn

Hock-Hai Teo
Department of Information Systems and Analytics
National University of Singapore
15 Computing Drive, Singapore, 117418
teohh@comp.nus.edu.sg

Abstract

The AI-generated voice (AIGV) has been widely applied in the short video industry to facilitate video creation. However, the impact of AIGV on digital consumer engagement (DCE) remains unclear. In light of this, the study investigates the effect of AIGV on DCE by analyzing a panel dataset with observations of 21,541 videos for 3,647 content creators on TikTok. Preliminary results of a series of fixed-effect panel regressions reveal that using AIGV has a significantly negative effect on DCE (5.4% reduction in the number of likes, 5.2% reduction in the number of comments, and 7.4% reduction in the number of shares). Our further analyses show that this negative effect is particularly significant at the rising action stage of short videos. With these findings, this study is expected to have theoretical contributions to the literature on short videos and practical implications about the appropriate usage of AIGV.

Keywords: AI-generated voice, digital user engagement, short video, AIGC

Introduction

The short video industry has experienced remarkable growth over the past few years, emerging as a major force in the digital landscape and captivating audiences worldwide. It is reported that the short video industry's market size reached 1.52 billion USD in 2022, with users spending an average of 95 minutes per day consuming video content (Grand View Research 2022). Most recently, the integration of AI-generated content (AIGC) into the short video industry has the potential to further expand this market. Various innovative AI applications, such as voice-overs, text-to-speech, and even content ideation for videos (Kumar et al. 2023), have been recognized for their ability to enhance the efficiency and cost-effectiveness of short video production (Kim et al. 2020). These advancements can contribute significantly to the ongoing boom in the short video industry.

While AIGC has become increasingly prevalent in the short video industry, it remains little known how AIGC may influence digital user engagement (DCE). DCE refers to consumers' interactions with a brand or business in a digital environment (Gavilanes et al. 2018). It not only reflects the popularity and reputation of content creators but also has a significant impact on platform revenue and sustainability (Santos et al. 2022). Some studies have shown that DCE may be facilitated by the generation of high-quality short videos with AIGC (Zhang 2023). However, researchers are also concerned about the potential negative influence of AIGC on audience perception, which may lead to a decline in DCE (Wang et al. 2022). Considering this paradox about the impact of AIGC on DCE, this study focuses on a key facet of AIGC in the short video industry: AI-generated voice (AIGV) (Habibi and Salim 2021). We aim to contribute to the existing literature by providing empirical evidence on how the use of AIGV influences DCE. To the best of our knowledge, prior research has only started exploring methods for distinguishing AIGV from human-generated voice based on voice signal similarity. Such methods, however, may not be universally applicable to the diverse range of AI voice generators available in the market, posing a significant methodological challenge in quantifying the effect of AIGV on DCE (Zhang 2023). To address these research gaps, this research aims to answer the following research question.

RQ: How will the usage of AI-generated voice affect digital consumer engagement in short videos?

Specifically, we collect a unique dataset from TikTok to empirically investigate the effect of AIGV on DCE in short videos. This dataset comprises 3,647 content creators and their 21,541 videos posted between December 1, 2019, and March 10, 2023. We employ state-of-art voice activity processing algorithms and a deepfake voice detection model to automatically detect the usage of AIGV in each video (Kawa et al. 2022). Given the potential self-selection issues, we estimate the effects coupled with propensity score matching using fixed effect panel regression. Our preliminary results reveal that the adoption of AIGV had a negative and significant influence on DCE (5.4% reduction in the number of likes, 5.2% reduction in the number of comments, and 7.4% reduction in the number of shares). Moreover, we examine the heterogeneity in the effect of AIGV across different stages of short videos, which deepens our understanding of the circumstances under which AIGV may enhance user engagement in short videos with finer granularity. Building upon Gustav Freytag's pyramid, we segment short videos into five stages, namely exposition, rising action, climax, falling action, and denouement (Quesenberry and Coolsen 2019). Our preliminary findings suggest that the adoption of AIGV at the rising action stage has a more significant negative effect on DCE, with a 6.4% reduction in the number of likes, comments, and shares. As this is an ongoing study, we will conduct further analyses following our plan described in the last section.

Our research has both theoretical contributions and practical implications. First, our research contributes to the growing body of literature on human-AI interaction by investigating the effect of AIGC on user engagement in the context of online content creation. To the best of our knowledge, this study is among the first to empirically investigate the potential effect of AIGV on DCE, a key aspect of content creation that has received scant attention in the existing literature. Second, our research contributes to the Information Systems (IS) literature on unstructured data analysis by using state-of-art deep learning models to automatically detect AIGC in audio. Unlike prior studies that focus on a limited set of specific AI voice generators (Zhang 2023), our methodology allows for a scalable and accurate identification of AIGV in large datasets. Finally, this study also offers valuable insights for content creators and businesses interested in adopting AIGC technology in their video production. For example, our results will help them understand the effect of AIGV on user engagement and determine when AIGV adoption would be advantageous.

Theoretical Background

AI-generated voice and short videos

The advancement of artificial intelligence technology has significantly streamlined workflows and enhanced work efficiency. Beyond merely offering textual guidance, AI has showcased its vast potential in multimodal processing in recent years. This cutting-edge development enables AI to transform text into various formats, including audio. Many short video platforms have embraced these techniques and developed official functions to support AIGV, including innovative applications such as voice-overs, text-to-speech, and voice content ideation (Kumar et al. 2023). As a result, the global AI voice generator market is projected to grow significantly from USD 1,210 million in 2022 to USD 4,889 million in 2032 (Market.us 2023).

Given the widespread use of AIGV in the short video industry, there is still a lack of research on how audiences will respond to videos that incorporate these generated voice (Kim et al. 2020). Researchers have examined the difference between AI-generated and human-generated voices on audiences' explicit and implicit perceptions (Wang et al. 2022). While explicit perception refers to the conscious perception of voice quality, fluency, and linguistic style, implicit perception involves changes in experience and thought, such as novelty, emotion evocation, and expression of empathy aroused by the voice (Kihlstrom et al. 1992) Although prior research has shown that AI-generated audio contents are comparable to the human-generated audio content in terms of effectiveness, perceived quality, readability, and credibility, they still lack the ability to convey implicit perception (Wang et al. 2022). Moreover, while using AIGV can significantly enhance video creation efficiency in the short term, it may negatively affect content creators' creative efforts in the long run (Zhang 2023). This decreased creative effort can further exacerbate the lack of implicit perception in AIGV, as most current AI techniques still depend on human guidance and input for generating new content. Hence, in this study, we aim to understand audience response to short videos utilizing AIGV. We hope to provide insightful guidance for content creators and platforms and facilitate the proper application of such AI technology.

Digital consumer engagement

DCE has been well known to reflect the audience's interactions with content creators. It is crucial for both short video content creators and platforms since it not only indicates creator popularity and reputation, but also correlates with platform revenue and sustainability (Santos et al. 2022). According to Gavilanes et al. (2018), there are four levels of DCE: neutral consumption, positive filtering, cognitive and affective processing, and advocacy. Neutral consumption is the lowest level of engagement, in which consumers primarily view short video content without actively engaging with the content. Positive filtering represents a moderate level of engagement and involves a more emotional investment from consumers. Cognitive and affective processing necessitates a higher level of DCE, as it demands more time and cognitive effort to engage. Advocacy, the highest level of DCE, entails a more robust cognitive and emotional investment, value co-creation, self-expression, and content dissemination (Schivinski et al. 2016).

Prior literature in the video industry has extensively addressed the importance of DCE to content creators. Neutral consumption serves as the foundation for higher-level engagement behaviors, and it can be indicated by video views (Munaro et al. 2021). During this stage, users can swipe down to skip the current video and start watching a new one without any prior selection, which reflects the minimal engagement and investment required from users (Kang and Lou 2022). In addition, positive filtering, reflected by users' initial emotional engagement with the content, is often signalled by video likes (Habibi and Salim 2021). In the context of short videos, voice characteristics that can arouse emotion, such as linguistic style, overall tone, and sense of humour, have been identified as crucial factors that influence users' positive filtering decisions (Munaro et al. 2021). Moreover, cognitive and affective processing allows users to express their thoughts, concerns, and feedback on the content, where video comments serve as the indicator. Consumers usually comment when they find the content meaningful (Schivinski et al. 2016). Therefore, voice factors that can evoke emotional resonance, including expressiveness, and emotional valence, have been identified as important factors for users when deciding to leave a comment (Habibi and Salim 2021). Finally, advocacy involves a more profound cognitive and emotional investment, self-expression, and content dissemination (Gavilanes et al. 2018). It is often indicated by video sharing, where users express their thoughts, feelings, or share valuable information. Video sharing implies that users perceive the content as highly valuable and worth recommending to others (Habibi and Salim 2021). Additionally, users may share videos to maintain social connections and exert social influence, where the novelty and uniqueness of the content are critical, as the shared video represents the individual sharing it (Song et al. 2023).

To summarize, different levels of DCE for short videos are shaped by different factors. First, positive filtering is influenced by voice characteristics that can arouse emotion. Second, cognitive and affective processing depends on voice factors that can evoke emotional resonance. Last, advocacy relies on the novelty and uniqueness of video content. In the next subsection, we analyze how the usage of AIGV may influence positive filtering, cognitive and affective processing, and advocacy of short videos.

Hypothesis Development

We draw upon the dual process theory (Groves and Thompson 1970) to examine the effect of AIGV on DCE. Dual process theory is a psychological framework that proposes two distinct cognitive processes to explain

human decision-making and information-processing. Specifically, System I process is primarily based on heuristic cues in the message (Groves and Thompson 1970). Considering that the average duration of short videos is approximately 30 seconds, users' capacity for deliberate thinking is often constrained. Therefore, people tend to make quick, heuristic decisions in a short time when viewing short videos, which is aligned with the System I process. Additionally, short video platforms have streamlined the process of liking videos to promote user engagement, with viewers expressing their appreciation by simply double-tapping the screen or clicking the like button. Such a rapid and automatic response depends more on the intuitive, reflexive, and less cognitively demanding reactions to the content (Ingold et al. 2018). Hence, in the context of short videos, elements such as linguistic style, prevailing tone, and sense of humour play a pivotal role in eliciting intuitive emotional reactions from viewers, and influencing their positive voice filtering decisions (Munaro et al. 2021). However, the deficiency in implicit perception makes it challenging for AIGV to customize such an overall tone and sense of humour. Therefore, we hypothesize that:

H1: The usage of AI-generated voice will negatively influence the positive filtering of short videos.

System II process, on the other hand, is a slower, more deliberate, and analytical process that requires conscious effort and rational thinking, which represents an increased level of elaboration of the message (Groves and Thompson 1970). We argue that cognitive and affective processing and advocacy are aligned with the System II process, which necessitates deliberate thought, reflection, and the expression of personal opinions (Gavilanes et al. 2018). As discussed earlier, AIGV frequently falls short when it comes to conveying thought and experience changes, which makes it less compelling for users to engage with the content through cognitive and affective processing (Wang et al. 2022). Moreover, the utilization of AIGV may pose challenges to fostering a positive creator-audience interaction. Due to the difficulty of customizing AIGV to meet the specific needs expressed in viewer feedback, audience members may feel discouraged from engaging in further discussions or providing feedback on subsequent videos (Habibi and Salim 2021). Based on this, we hypothesize that:

H2: The usage of AI-generated voice will negatively influence the cognitive and affective processing of short videos.

Advocacy involves a more profound cognitive and emotional investment, value co-creation, self-expression, and content dissemination (Gavilanes et al. 2018). Apart from the shortcomings in implicit perception and building social interaction, videos that utilize AIGV often exhibit similar voice features. For example, AIGV from current text-to-speech functions tend to have comparable voice tones (Zhang 2023). This will be further exacerbated by the short video platform recommendation mechanism, which often provides similar video recommendations (Kang and Lou 2022). The increased similarity of recommended videos will further increase the sense of familiarity and diminish the perceived uniqueness of the videos, negatively influencing advocacy. Consequently, we hypothesize that:

H3: The usage of AI-generated voice will negatively influence the advocacy of short videos.

Research Context and Data

Research context. We choose TikTok as our research context for several reasons. To begin with, the popularity of TikTok among users provides us with an abundant number of short videos with diverse content, which ensures the generalizability of our preliminary findings. As of 2021, TikTok had an estimated 800 million monthly active users (Habibi and Salim 2021), making it one of the most popular short video platforms. These active users on TikTok are also allowed to share a wide range of content, including their records of singing, dancing, cooking, or performing various other activities. Therefore, with a great number of active users and the freedom of expression on TikTok, we can access enough diverse videos for our empirical investigation. In addition, the diversity enables us to make an empirical comparison between videos with AIGV and those with human-generated voice, thus drawing conclusive results about the impact of AIGV on DCE on short video platforms.

Data collection. Relying on the API provided by TikTok, we start our data collection process. Initially, we randomly sample over 5,000 content creators and collect metadata for each creator, including their number of followers, number of hearts, and number of likes. While the number of hearts typically refers to the number of times a creator has been liked by users, the number of likes refers to the total number of likes received in all the creator's videos. Both metrics are commonly used to gauge the creator popularity on the

platform. For each creator, we also gather the metadata of their latest 5 to 10 videos and download the raw video files. Subsequently, we extract a number of features of each video, including the description, hashtags, and duration of each video. Consistent with prior literature (Munaro et al. 2021), we also record the number of likes, comments, and shares for each video as indicators of positive filtering, cognitive and affective processing and advocacy respectively, which are our dependent variables. In Table 1, we list the variables that we construct based on the data collected.

Variable	Description	Mean	Std.	Min	Max
LikeCount_{ij}	Number of likes of video_{ij} (log)	6.690	3.118	0.000	16.739
CommentCount_{ij}	Number of comments of video_{ij} (log)	2.732	2.246	0.000	13.177
ShareCount_{ij}	Number of shares of video_{ij} (log)	2.739	2.438	0.000	13.496
$\text{AI\_voice\_usage}_{ij}$	Whether video_{ij} uses AI-generated voice	0.505	0.500	0.000	1.000
$\text{CreatorFollower}_{i}$	Number of followers of creator_{i} (log)	10.719	2.618	2.708	17.876
CreatorHeart_{i}	Number of hearts creator_{i} received (log)	13.405	3.109	3.784	21.193
CreatorLike_{i}	Number of likes creator_{i} received (log)	6.574	2.695	0.000	12.937
CreatorVideo_{i}	Total number of videos of creator_{i} (log)	5.059	1.297	0.693	8.894
$\text{VideoDuration}_{ij}$	Duration of the video_{ij} (second, log)	3.579	1.031	0.693	6.397
$\text{VoicePercentage}_{ij}$	Percentage of the video's duration that is covered by the AI-generated voice	0.688	0.269	0.008	1.000
Hashtag_{ij}	The number of hashtags of video_{ij} (log)	1.559	0.825	0.000	4.575
Sentiment_{ij}	Sentiment score for video_{ij}	0.127	0.284	-1.000	1.000
Category_{ij}	Classification of the video_{ij} content	7.700	4.251	1.000	16.000
$\text{ReleaseGT\_7days}_{ij}$	Whether the upload date of video_{ij} is more than 7 days from the fetching date	0.942	0.234	0.000	1.000
ReleaseTime_{ij}	The time gap between video_{ij} upload date and fetching date (unit: days, log)	3.660	1.067	0.474	7.087

Table 1. Variable Description and Summary of Statistics

Variable construction. We constructed three key variables using machine learning models, i.e., $\text{AI\_voice\_usage}_{ij}$, Category_{ij}, and Sentiment_{ij}. $\text{AI\_voice\_usage}_{ij}$ is our key independent variable, which is a binary variable indicating whether video *j* created by content creator *i* uses AIGV or not. To construct this variable, we distinguish videos with AIGV from those with human-generated voice following a three-step procedure. In the first step, we identify the language of the videos collected and filter out non-English videos to mitigate language-related confounding factors and ensure the accuracy of AI_voice_usage. We retrieve the videos from TikTok API and employ DeepL Translate to identify and retain only English videos. In the second step, we employ several state-of-the-art audio analysis algorithms to extract voice information from massive audio data that is cluttered with background music, noise, and silence. First, we use Fast Forward moving picture experts group (FFmpeg) to separate audio from videos and apply a pre-trained model called Spleeter to separate and remove background music. Next, we use the pyannote library to detect and annotate speaker voice activity within the videos. The library annotates the start and end times of every instance of human voice activity, enabling us to exclude segments containing silence or noise. In the final step, we apply a machine learning model to distinguish between human-generated voice and AIGV. Previous literature primarily focuses on the similarity between voice signals, using identified AIGV as a benchmark (Zhang 2023). However, this method is only effective for specific AI tools, and it is difficult to generalize, particularly when there are many types of AI voice tools available. To address this challenge, we train a DeepFake voice detection model (LCNN with adversarial training; Kawa et al. 2022) to examine the usage of AI voice based on the ASVspoof 2021 Deepfake dataset (Liu et al. 2022). In addition to the traditional LCNN detection architectures, we employ adversarial training performed by adaptive training methods to enhance the model's robustness against adversarial attacks (Kawa et al. 2022). Our model

achieved an accuracy of 98.05% on a 250 million-sample testing dataset. After these steps, our final sample comprises 3,647 creators and 21,541 videos.

$Category_{ij}$ and $Sentiment_{ij}$ are two control variables constructed to mitigate the confounding effect of video content. $Category_{ij}$ indicates the video type for video j created by creator i. We classify videos into different categories based on their content. As TikTok does not have official categories for videos, we refer to video categories used in YouTube Shorts to pre-define a classification of video content, which includes 16 different types: automotive, comedy, education, entertainment, film, gaming, DIY, music, news, pets, science, sports, travel, vlogs, non-profit, and others. Next, we utilize the spaCy library to calculate text similarity between the hashtags of the videos and our predefined categories. The video's category is then determined by the category with the highest similarity score. As for $Sentiment_{ij}$, it quantifies the sentiment of video j created by creator i using the TextBlob library, which assigns a score to video j ranging from -1 (most negative) to 1 (most positive). This sentiment score can capture the overall emotional tone of the video.

Variable	Before matching			After matching			t-test
	(N_{tr} =10,876, N_{co}=10,665)			(N_{tr} =10,876, N_{co}=5,773)			p-value
	$Mean_{Treated}$	$Mean_{Control}$	Std.	$Mean_{Treated}$	$Mean_{Control}$	Std.	
$CreatorFollower_i$	10.780	10.657	4.7	10.780	10.764	0.6	0.664
$CreatorHeart_i$	13.461	13.348	3.6	13.461	13.432	0.9	0.496
$CreatorLike_i$	6.661	6.485	6.5	6.661	6.622	1.4	0.635
$CreatorVideo_i$	5.013	5.104	-7.0	5.013	5.012	0.1	0.953
$VideoDuration_{ij}$	3.463	3.698	-22.9	3.463	3.453	1.0	0.481
$VoicePercentage_{ij}$	0.689	0.688	0.4	0.689	0.686	1.0	0.474
$Hashtag_{ij}$	1.554	1.562	-1.0	1.554	1.546	0.9	0.484
$Sentiment_{ij}$	0.127	0.126	0.7	0.127	0.130	-0.9	0.495
$ReleaseGT\_7days_{ij}$	0.944	0.939	2.1	0.944	0.942	0.7	0.617
$ReleaseTime_{ij}$	3.675	3.644	2.9	3.675	3.680	-0.4	0.745

Table 2. Covariate Balance Before and After Matching

Estimation sample. After data collection and construction, we implement propensity score matching (PSM) to ensure the comparability of treatment and control groups. PSM is used to control for potential confounding effects of unobserved covariates. In this study, as factors like prior AI experience and video production difficulty may influence content creators' adoption of AIGV and further make it difficult to isolate the true effect of AIGV on DCE, we need to use PSM to minimize the effects of confounding factors. We first use all the video-related variables (X_{ij}) and some observed creator-related variables ($Author_i$) to generate the propensity scores. Next, we employ the single nearest-neighbor function with a value of one without replacement for 1-on-1 matching. It results in 10,876 treated videos and 5,773 control videos, dropping 4,892 units in the control group. Finally, our estimation sample includes 16,649 videos for 3,547 content creators. In Table 2, we display the covariate balance before and after the PSM.

Empirical Model Specification

To investigate the effect of AIGV on DCE, we specify a panel regression model as follows.

$$DCE_{ij} = \beta_0 + \beta_1 * AI\_voice\_usage_{ij} + \beta_2 * X_{ij} + \gamma_i + \varepsilon_{ij}$$

Where DCE_{ij} represents the user engagement situation for video j created by content creator i. X_{ij} indicates the vector of control variables for $video_{ij}$. γ_i denotes the creator-specific fixed effect and ε_{ij} is the residual error term. Moreover, to address the skewed distribution of continuous variables, we apply a logarithmic transformation to all continuous variables in our analysis to allow for elasticity interpretation.

We further examine the effect of the usage of AIGV on DCE in the different stages of short videos. Drawing from Gustav Freytag's pyramid, we segment the videos equally into five stages based on their durations, including exposition, rising action, climax, falling action, and denouement (Quesenberry and Coolsen 2019). Then, we identify the usage of AIGV in each stage using machine learning methods (AI_voice_stage). Based on this setup, the utility of model can be stated as:

$$DCE_{ij} = \beta_0 + \sum_{k=1}^{5} \beta_{1k} * AI\_voice\_stage_{ijk} + \beta_2 * X_{ij} + \gamma_i + \varepsilon_{ij}$$

Where $AI\_voice\_stage_{ijk}$ represents usage of AIGV in stage k of video j created by content creator i.

Preliminary Results

Variable	LikeCount	CommentCount	ShareCount
AI_voice_usage	-0.053*	-0.051*	-0.072**
	(0.030)	(0.027)	(0.031)
VideoDuration	0.291***	0.277***	0.356***
	(0.026)	(0.023)	(0.026)
VoicePercentage	-0.139**	-0.121**	-0.126*
	(0.070)	(0.060)	(0.073)
Hashtag	0.105**	0.040	0.135***
	(0.046)	(0.041)	(0.047)
Sentiment	-0.069	-0.084*	-0.068
	(0.054)	(0.046)	(0.054)
ReleaseTime	-0.224**	-0.104	-0.098
	(0.095)	(0.079)	(0.087)
ReleaseGT_7days	0.763***	0.536***	0.662***
	(0.177)	(0.155)	(0.155)
Constant	5.728***	1.634***	1.123***
	(0.377)	(0.327)	(0.365)
Observations	16,649	16,649	16,649
R-squared	0.023	0.024	0.029

Table 3. Estimation result of fixed effect regression

*Note: Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1.*

Table 3 presents our preliminary results from our fixed effect model estimation[1]. The usage of AIGV in TikTok videos results in a 5.4% reduction in the number of likes (p = 0.057), a 5.2% reduction in the number of comments (p = 0.076), and a 7.4% reduction in the number of shares (p = 0.021), supporting H1, H2, and H3 respectively. Overall, our preliminary results suggest that the adoption of AIGV has a significantly negative effect on all kinds of DCE.

[1] We conduct Hausman test among three models to help determine the appropriate specification (fixed effect model vs. random effect model). Results validate our model specification (like model, chi2(22) = 65.13, p < 0.001; comment model, chi2(22) = 83.38, p < 0.001; share model, chi2(22) = 105.04, p < 0.001).

Variable	LikeCount	CommentCount	ShareCount
AI_voice_stage Exposition	-0.026	0.001	-0.038
	(0.034)	(0.029)	(0.035)
AI_voice_stage Rising Action	-0.062*	-0.062**	-0.062*
	(0.034)	(0.030)	(0.034)
AI_voice_stage Climax	-0.006	0.009	-0.022
	(0.034)	(0.030)	(0.035)
AI_voice_stage Falling Action	0.025	0.005	0.036
	(0.034)	(0.030)	(0.035)
AI_voice_stage Denouement	0.018	-0.007	0.001
	(0.035)	(0.031)	(0.035)
Observations	16,649	16,649	16,649
R-squared	0.023	0.024	0.029

Table 4. Estimation results of usage of AI-generated voice in different stages

*Note: Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1.*

Table 4 reveals that the usage of AIGV at the rising action stage has a significant and negative effect on all three dimensions of DCE. Specifically, the impact on the number of comments is statistically significant at the 5% level (p = 0.040). This suggests that the use of AIGV might impede the audience's deliberate and analytical processes, thereby having a more pronounced negative impact on their cognitive and affective decision-making. Our preliminary analyses conclude that content creators are better off using AIGV towards the middle or end of the video if they aim to engage users.

Conclusion and Future Work

This study empirically examines the effect of AIGV adoption on DCE on short video platforms. Our preliminary results reveal a negative association between the adoption of AIGV and user engagement on short video platforms. More importantly, we find that the effect of AIGV is particularly pronounced at the rising action stage, while its influence diminishes in the middle or end of the video. With these preliminary results, we expect to raise content creators' caution to use AIGV at the rising action stage and suggest strategic placement of such content throughout the video to mitigate the potential negative impact of AIGV.

In the future, there are several additional works that we aim to complete. First, we plan to collect more longitudinal data to gain a deeper understanding of the effect of AIGV on user engagement over time. Second, we will conduct additional subgroup analyses to examine the heterogeneity in the impact of the usage of AIGV. For example, variables such as video type and the content creator's number of followers can be used to classify videos into different groups. Third, we will explore the influence of AIGV characteristics in detail, such as its linguistic style, intonation and vocal bursts. Finally, we will conduct additional robustness checks to further validate our empirical findings. Currently, we have utilized the Heckman two-stage model to address selection bias arising from content creators' prior experience with AIGV. The results of this check are consistent with our main findings.

Acknowledgements

This research / project is supported by the National University of Singapore, Singapore under its Provost's Chair Award, Humanities and Social Sciences Seed Fund, and the National Natural Science Foundation of China (Grant Numbers: 72301279).

References

Gavilanes, J. M., Flatten, T. C., and Brettel, M. 2018. "Content Strategies for Digital Consumer Engagement in Social Networks: Why Advertising Is an Antecedent of Engagement," *Journal of Advertising* (47:1), Routledge, pp. 4–23.

Grand View Research. 2022. "Short Video Platforms Market Size and Share Report," *Short video platforms market size & share report, 2030*.

Groves, P. M., and Thompson, R. F. 1970. "Habituation: A Dual-Process Theory," *Psychological Review* (77), US: American Psychological Association, pp. 419–450.

Habibi, S. A., and Salim, L. 2021. "Static vs. Dynamic Methods of Delivery for Science Communication: A Critical Analysis of User Engagement with Science on Social Media," *PLOS ONE* (16:3), Public Library of Science, p. e0248507.

Ingold, P. V., Dönni, M., and Lievens, F. 2018. "A Dual-Process Theory Perspective to Better Understand Judgments in Assessment Centers: The Role of Initial Impressions for Dimension Ratings and Validity," *Journal of Applied Psychology* (103), US: American Psychological Association, pp. 1367–1378.

Kang, H., and Lou, C. 2022. "AI Agency vs. Human Agency: Understanding Human–AI Interactions on TikTok and Their Implications for User Engagement," *Journal of Computer-Mediated Communication* (27:5), (L. Humphreys, ed.), p. zmac014.

Kawa, P., Plata, M., and Syga, P. 2022. *Defense Against Adversarial Attacks on Audio DeepFake Detection*, arXiv.

Kihlstrom, J. F., Barnhardt, T. M., and Tataryn, D. J. 1992. "The Psychological Unconscious: Found, Lost, and Regained," *American Psychologist* (47), US: American Psychological Association, pp. 788–791.

Kim, J., Shin, S., Bae, K., Oh, S., Park, E., and del Pobil, A. P. 2020. "Can AI Be a Content Generator? Effects of Content Generators and Information Delivery Methods on the Psychology of Content Consumers," *Telematics and Informatics* (55), p. 101452.

Kumar, Y., Koul, A., and Singh, C. 2023. "A Deep Learning Approaches in Text-to-Speech System: A Systematic Review and Recent Research Perspective," *Multimedia Tools and Applications* (82:10), pp. 15171–15197.

Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H., Kinnunen, T., Todisco, M., Yamagishi, J., Evans, N., Nautsch, A., and Lee, K. A. 2022. *ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild*, arXiv.

Market.us. 2023. "Ai Voice Generator Market Share, trends, analysis, forecast 2032.".

Munaro, A. C., Hübner Barcelos, R., Francisco Maffezzolli, E. C., Santos Rodrigues, J. P., and Cabrera Paraiso, E. 2021. "To Engage or Not Engage? The Features of Video Content on YouTube Affecting Digital Consumer Engagement," *Journal of Consumer Behaviour* (20:5), pp. 1336–1352.

Petty, R. E., and Cacioppo, J. T. 1986. "The Elaboration Likelihood Model of Persuasion," in *Advances in Experimental Social Psychology* (Vol. 19), L. Berkowitz (ed.), Academic Press, pp. 123–205.

Quesenberry, K. A., and Coolsen, M. K. 2019. "Drama Goes Viral: Effects of Story Development on Shares and Views of Online Advertising Videos," *Journal of Interactive Marketing* (48), pp. 1–16.

Santos, Z. R., Cheung, C. M. K., Coelho, P. S., and Rita, P. 2022. "Consumer Engagement in Social Media Brand Communities: A Literature Review," *International Journal of Information Management* (63), p. 102457.

Schivinski, B., Christodoulides, G., and Dabrowski, D. 2016. "Measuring Consumers' Engagement With Brand-Related Social-Media Content: Development and Validation of a Scale That Identifies Levels of Social-Media Engagement with Brands," *Journal of Advertising Research* (56:1), Journal of Advertising Research, pp. 64–80.

Song, H., So, J., Shim, M., Kim, J., Kim, E., and Lee, K. 2023. "What Message Features Influence the Intention to Share Misinformation about COVID-19 on Social Media? The Role of Efficacy and Novelty," *Computers in Human Behavior* (138), p. 107439.

Wang, J., Li, S., Xue, K., and Chen, L. 2022. "What Is the Competence Boundary of Algorithms? An Institutional Perspective on AI-Based Video Generation," *Displays* (73), p. 102240.

Zhang, X. 2023. "How does AI-Generated voice affect online video creation? Evidence from TikTok", *[MA thesis]*, University of British Columbia.