Luo, Y.M., Ling, Z.Y., & Meng, L.Y. (2025). How acoustic cues influence user participation in travel Videos? In Li, E.Y. *et al.* (Eds.) *Proceedings of The International Conference on Electronic Business, Volume 25* (pp. 206-214). ICEB'25, Hanoi, Vietnam, August 21-25, 2025

How Acoustic Cues Influence User Participation in Travel Videos?

Yumei Luo [1]
Zeyun Ling [2]
Luoyan Meng [3, *]

*Corresponding author
[1] Yunnan University, Kunming, Yunnan, China, luoyumei@ynu.edu.cn
[2] Yunnan University, Kunming, Yunnan, China, 3044622850@qq.com
[3] Yunnan University, Kunming, Yunnan, China, 1803581682@qq.com

ABSTRACT

As a new format for promoting tourism and culture, travel vlogs are rapidly evolving into a pivotal medium for destination image marketing and a primary channel for tourist information acquisition, establishing novel paradigms for information sharing and activity planning in tourism. Existing research predominantly focuses on destination image construction/perception, tourist behavioral intentions, attraction marketing strategies, and communication effectiveness, while neglecting audio feature analysis. Addressing this gap, we employ face as a moderator, audio features as independent variables (IVs), and audience Participation as the dependent variable (DV) to examine how acoustic signals influence Danmaku comment participation. Results reveal that audio features interact with faces presence to co-shape Participation. This study elucidates the critical roles of acoustic signals and face in travel Vlog impact dynamics and provides actionable insights for creators to strategically leverage audio-visual synergies.

Keywords: Travel vlog videos, background sounds, host voices, Danmaku, participation behavior.

INTRODUCTION

With the popularization of mobile Internet and the explosive growth of short video platforms, short videos have emerged as a significant medium for users' daily entertainment and information acquisition. According to the "China Online Audio-Visual Development Research Report 2025," by the end of 2024, the number of short-video users in China had reached 1.04 billion, with an average daily usage time exceeding 2.5 hours. As one of the prominent areas within short videos, the creation volume and demand for travel vlogs are increasing year by year. However, audience attention towards extensive travel vlogs is polarized. High-quality travel vlogs can effectively engage viewers, prompting them to "stop and pay attention" and become fully immersed in the content. In contrast, low-quality or homogeneous travel vlogs tend to elicit only inefficient interactions characterized by "swiping away and forgetting." Consequently, attracting audience attention has emerged as a focal concern among creators.

As one of the critical factors influencing audience participation, sound elements possess the ability to evoke emotions through their unique sensory stimulation. This enhances the sense of presence and reality associated with "online tourism," thereby capturing viewer interest and promoting behaviors such as browsing, collecting, sharing, commenting, and other forms of participation. Nevertheless, despite being an "invisible narrator" within travel vlog videos, the significance of sound elements is frequently overlooked by creators. Currently, there is a predominant focus on the direct impact of visual components on audiences (Luo *et al.*, 2025) while the subtle influence exerted by auditory elements—such as host voices and background music—remains largely unacknowledged.

Issues such as selecting voices based on emotional resonance, mismatching background sounds with visuals, and reliance on templated host voices have consistently posed challenges for creators during production. Furthermore, existing research on short videos has paid scant attention to understanding both the mechanisms at play regarding sound elements in travel vlog videos as well as their resultant effects. Therefore, this paper aims to investigate sound elements within travel vlog videos as its primary research subject—a pursuit that holds both practical implications and theoretical significance.

LITERATURE REVIEW

Sound Elements

On short video platforms, the proliferation of travel vlogs has significantly diminished audience patience. Capturing the audience's attention remains a critical concern for creators. According to the "golden three-second rule" in the realm of short videos, enhancing video completion rates, interaction rates, and platform recommendation volumes necessitates that creators engage viewers within the first three seconds of their content. Sound serves as an "emotional guide" in travel vlogs and can elicit emotional responses from audiences more rapidly than visual elements (Hazmoune & Bougamouza, 2024), thereby swiftly capturing active attention (Simmonds *et al.*, 2020). Rutten *et al.*(2019). posited that human brains exhibit heightened sensitivity to auditory stimuli compared to visual stimuli. Sound has the capacity to activate multiple brain regions, quickly drawing personal attention (Charest *et al.*, 2009) and facilitating accelerated cognitive processing (Aeschlimann *et al.*, 2008). Consequently,

investigating the mechanisms by which sound elements function in short videos is particularly important. In travel vlog productions, sound elements primarily encompass background sounds and host voices. Background sounds refer to ambient noises, natural sounds, human sound effects, or music present in the video that are not produced by the host; whereas host voice denotes the primary auditory medium through which hosts convey travel information, narratives, and emotions via oral expression. These voices capture the blogger's genuine emotions during their travels, creating a more comprehensive and authentic experience for viewers and enhancing their perceptions, attitudes, emotions and behaviours towards the video.

The purpose of background sound is to recreate the auditory landscape of a travel destination and assist the audience in constructing an alternative experience. Its key acoustic features include spectral centroid (SpeCentroid) and zero-crossing rate (ZCR). The spectral centroid represents the center of gravity of spectral energy, influencing the perceived "brightness or softness" of sound. Variations in the height of the spectral centroid significantly affect video users' auditory experiences. Sounds characterized by high spectral centroids are perceived as brighter (e.g., birdsong), while those with low spectral centroids are experienced as softer (e.g., wind). Furthermore, the spectral centroid is closely linked to emotional arousal among audiences; higher values correspond to more pleasant emotions conveyed by background music (Přibil & Přibilová, 2011), which can evoke associations with "pleasure and relaxation." Conversely, lower values tend to convey milder emotions, facilitating a "heavy and healing" atmosphere. Additionally, variations in the height of the spectral centroid directly influence audience participation behavior. Research indicates that when the average spectral centroid of background music increases by 10%, users' median viewing time extends by 23 seconds (Lin, Li & Mou, 2024). However, excessively high values may lead to feelings of "auditory discomfort," resulting in decreased completion rates (Fu *et al.*, 2024). Optimal user interaction occurs when the spectral centroid resides within a specific range; exceeding this range can induce perceptual fatigue (Swati Shilaskar *et al.*, 2023).

The zero-crossing rate quantifies how frequently a voice signal's amplitude crosses zero within a specified time interval, thereby determining sound complexity. A high zero-crossing rate corresponds to rich sound elements manifested through rapidly changing sounds; conversely, a low zero-crossing rate reflects simpler sound elements characterized by slowly changing sounds.The zero-crossing rate is intricately linked to the stability of audience attention. A moderate zero-crossing rate can effectively enhance video content, capture viewers' interest, increase the realism of the presentation, and improve audience participation behaviors. Conversely, an excessively high zero-crossing rate may distract viewers from the host's commentary, while a too low zero-crossing rate could result in a monotonous viewing experience (Gray & Markel, 1976).

The primary function of the host’s voice is to establish trust and evoke emotional resonance (Wang & Li, 2024). Key acoustic features that contribute to this include pitch and loudness. Pitch determines the height or lowness of the voice and is closely associated with emotional expression (Wang & Li, 2024). High pitch conveys liveliness and enthusiasm, whereas low pitch communicates calmness and reliability. Emotion plays a pivotal role in influencing user participation (Wang & Li, 2024). In short videos focused on climate communication, an increase in the host’s vocal pitch can significantly impact audience perceptions, thereby affecting their level of participation. In hotel videos, there is an inverted U-shaped relationship between pitch and audience participation. Moderate pitch can increase audience engagement behavior, while excessively increasing high pitch can make the audience feel uncomfortable and reduce their participation (Li *et al.*, 2025).

Loudness influences sound intensity directly impacting attention arousal and information reception efficiency; thus it affects audience participation behavior. Moderate loudness can swiftly attract viewer attention while maintaining focus—this approach has been found to substantially increase likes on videos and enhance viewer retention rates (Flores-Saviaga *et al.*, 2020). On the other hand, overly high loudness levels may induce auditory fatigue that negatively impacts completion rates. (Fu *et al.*, 2024) hereas insufficient loudness might diminish retention rates for travel tips due to "information ambiguity" (Gaver, 1993).

In summary, sound elements such as spectral centroid, zero-crossing rate, pitch, and loudness in short videos can significantly influence audience participation behaviors, including watching, liking, and commenting. Therefore, it is feasible to investigate audience participation behavior in travel vlog videos from the perspective of these auditory elements.

Audience Participation Behavior

Audience participation in travel vlog videos primarily refers to the interactive behaviors exhibited by viewers based on the content of the video. In this study, the number of Danmaku comments is employed as a quantitative measure of audience participation. In travel vlog videos, on one hand, the volume of Danmaku comments typically serves as a direct indicator of how actively audiences interact with specific video segments. By analyzing the count of Danmaku comments, we can assess viewer participation levels within short video formats. On the other hand, Danmaku comments are characterized by their visibility and shareability; they foster interactivity and provide entertainment value. These comments directly reflect audience reactions and attitudes toward the video content while assisting creators in enhancing video quality (Bai *et al.*, 2021). This improvement can further boost both viewership popularity and overall audience participation. Additionally, research has demonstrated that Danmaku comments not only signify viewer participation (Ni & Coupe, 2023) but also enhance attention (Shi, Zhao & Zhao, 2024), foster loyalty (Bai, Hu & Ge, 2019), and encourage consumption behaviors such as gifting or purchasing products related to the content presented. Consequently, utilizing Danmaku comment counts serves as a representative metric for quantifying audience participation.

RESEARCH MODEL AND HYPOTHESIS

The spectral centroid plays a pivotal role in determining the brightness or softness of background sound, while the zero-crossing rate influences the complexity of background sound. Additionally, pitch is responsible for conveying the height or lowness of host voices, and loudness determines its intensity. These nonverbal cues present in both background sounds and the host voice are crucial for effectively conveying information, mobilizing emotions (Wang & Li, 2024), and enhancing audience participation behavior. Sound elements can stimulate the audience's auditory perception, rich sound elements can intuitively reflect genuine feelings within travel vlog videos, thereby creating a more immersive and realistic experience for viewers. This enhancement improves overall attractiveness and captures audience attention towards travel vlogs (Flores-Saviaga *et al.*, 2020). Furthermore, it positively influences viewers' perceptions, attitudes, and emotions regarding the video content while promoting audience active participation, such as audience viewing or commenting.

On one hand, the spectrum centroid in the background sound of short videos, along with the pitch of the host's voice, conveys emotional cues to viewers. Within a moderate range, higher values of both spectrum centroid and pitch are associated with more positive and pleasant emotions (Přibil & Přibilová, 2011). Conversely, lower values correspond with less positive emotional expressions. Positive emotions elicited from travel vlog videos often resonate with audiences, this resonance encourages them to engage actively through Danmaku comments interactions. Thus, both the spectral centroid and pitch exert a beneficial influence on audience participation behavior.

On the other hand, factors such as zero-crossing rate in background sound of short videos and the loudness of the host's voice serve to capture audience attention effectively. Within reasonable limits, improved zero-crossing rate or increased loudness correlates with heightened audience focus (Flores-Saviaga *et al.*, 2020). The lower the zero-crossing rate or the smaller the loudness, the more likely it is that audience attention will diminish. The level of attention that viewers pay to a video is essential for facilitating participatory behaviors, thus, greater attentiveness makes it easier for audiences to engage actively Danmaku comments interactions. Therefore, both the zero-crossing rate and loudness positively influence audience participation behavior. In summary, this study puts forth the following hypotheses:

H1a: The loudness of the host's voice positively affects audience participation behavior;
H1b: The tone of the host's voice positively affects audience participation behavior;
H2a: The spectral centroid (SpeCentroid) of background sound positively affects audience participation behavior;
H2b: The zero-crossing rate (ZCR) of background sound positively affects audience participation behavior.

The human face is the most biologically and socially significant visual stimulus, with a unique ability to attract attention, convey information and effectively express emotions and intentions (Guido *et al.*, 2019). The neuroscience community generally posits that face perception is the most advanced visual skill among humans. In comparison to other sensory stimuli, faces elicit emotional responses more effectively and can be captured rapidly (Cerf, Frady & Koch, 2009). As a crucial social cue, human face is able to reconstruct the degree to which sound elements influence audience participation behaviours in short videos through stronger attraction. The presence of faces in travel vlogs can harmonise verbal communication, reduce cognitive load and facilitate closer, more efficient communication and interaction (Takeuchi & Nagao, 1993), Therefore, the presence of human faces in the video clearly and effectively conveys the content, increasing the audience's awareness and attention, and triggering their interactive behaviour. Consequently, this study proposes the following hypotheses:

H3a: The presence of human faces strengthens the effect of the host’s voice loudness on participation behavior;
H3b: The presence of a human face enhances the impact of the host's voice pitch on participation behavior;
H3c: The presence of human faces amplifies the effect of the spectral centroid (SpeCentroid) of background sounds on participation behavior;
H3d: The presence of human faces reinforces the effect of the zero-crossing rate (ZCR) of background sounds on participation behavior.

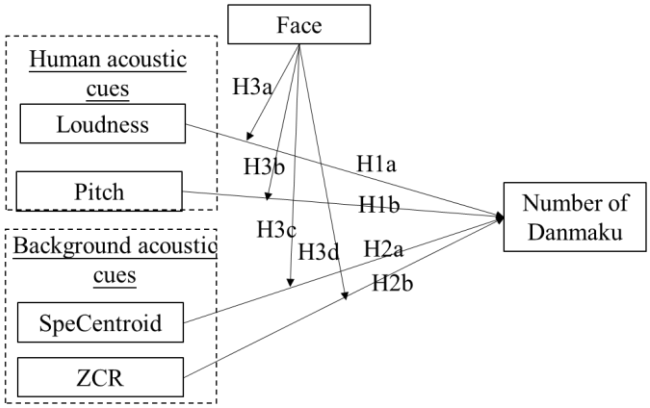

Figure 1： Research Model

DATA AND METHODOLOGY

Data Collection and Processing

The source of the video dataset is Bilibili (https://www.bilibili.com), which is considered to be the leading user-generated streaming video platform in China. The reason for choosing Bilibili is based on 2 main points: (1) Bilibili has a large user base，and its main audience are young people. Young people are more culturally sensitive and more active on the Internet, and are able to capture and reflect the trends of the times. (2) Unlike platforms such as Douyin and Kuaishou that feature extensive live-streaming commerce and promotional short videos, Bilibili hosts a greater volume of professionally created video content uploaded by users. These contents span diverse categories with higher specialization levels, encompassing the vast majority of travel vlogs among other video formats.

First, we set the search terms “travel” and “vlog” as the key search terms in the search bar on the homepage of Bilibili (see Figure 2). Excluded are videos published by official marketing accounts of tourist destinations (e.g., “China Tourism Recommendation Officer”) and videos in which the blogger is not recounting his or her own travel experiences. Second, Wood (2019) proposes that the recommended mean duration for vlogs under ideal conditions is 9 minutes and 43 seconds. Consequently, this study constrained selected travel vlogs to durations under 10 minutes while requiring Danmaku comments to exceed 30, thereby guaranteeing sufficient Danmaku-based interactive participation. According to the aforementioned criteria, 210 eligible travel vlogs were identified as of October 26, 2024, spanning diverse content categories including natural landscapes, historical culture, and travel guides.

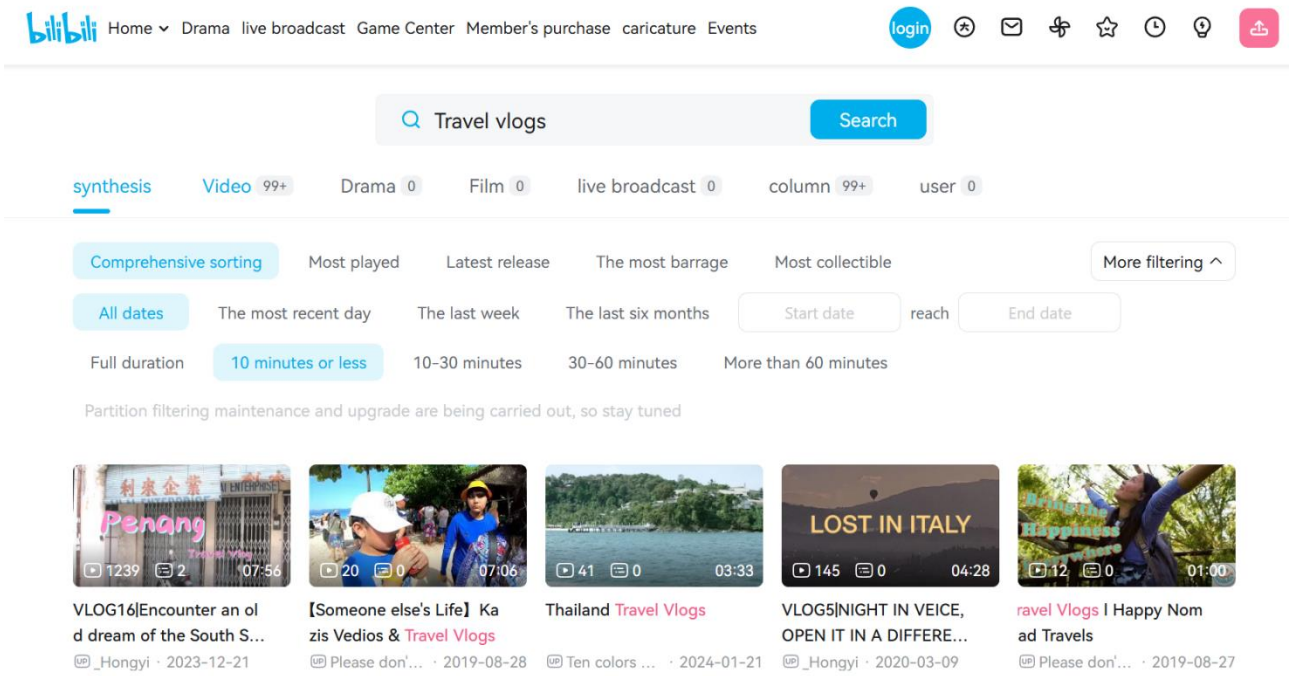

Figure 2: Filtering Travel Vlogs

Moreover, the low barriers to posting Danmaku comments may lead some viewers to post comments irrelevant to the video content or even non-compliant remarks. Therefore, in order to uncover the degree of correlation between travel vlog audio features and audience participation, this study needs to extract video clips that reflect the consistent views of most viewers. This study selects highlight segments as the research sample. According to Sun *et al.* (2024), the number of Danmaku comments represents user participation and the importance or attractiveness of the video, which can be used to determine the identification of highly involving sessions. Dai *et al.* (2024) argued that videos with more highly involved sessions have a greater amount of comment interactions and richer forms of interactions and that Danmaku comments are an important tool to reflect the characteristics of the video and the highly involved sessions. Following Xu *et al.* (2021), the number of Danmaku comments per second for each vlog is calculated, and if the number of the i-th second meets the criteria for the determination of a highlight moment (see Table 1), the i-th second of the video is identified as a highly involving moment. Thereafter, a 2-s video clip before the i-th second was selected as highly involving sessions. The acoustic features of the highly involving sessions were extracted as the independent variable in this study.

Table 1： Methods for identifying the highly involving moment

Description of symbols	
N_i	Number of pop-up comments in the i-th second of the video
$\bar{N}$	The average number of Danmaku comments per second
σ	The standard deviation of Danmaku comments by seconds
Criteria for determining highly involving moment： $N_i \geq \bar{N} + 2\sigma$	

The existing research suggests that video-related comments persisted for up to 5 seconds after the highly involving sessions (He & Tang, 2024). Therefore, in this study, the number of Danmaku comments within 5 seconds after highly involving moments was selected to measure viewer participation.

Through the aforementioned methodology, this study identified 1764 highly involving sessions from the 210 videos. Utilizing Python's moviepy library, we executed VideoFileClip function calls to extract and export video segments from the selected 210 travel vlogs according to specified timestamps. There are 609 samples with human faces present and 1155 samples with no faces present. This process established a solid data foundation for subsequent audio feature extraction and analysis of audience participation.

Feature Extraction

For human voices, we primarily extracted pitch and loudness features of travel vloggers' vocals. Loudness variability was measured by calculating the average standard deviation of the vlogger's sound amplitude in published videos; Pitch reveals the fundamental frequency of speech and determines the sound quality of the voice. The following three-step procedure was implemented to extract required acoustic signals:

(1) Employed Python's Moviepy package to extract audio data from video segments;

(2) Executed Spleeter package calls to separate vlogger vocals from background accompaniment, minimizing interference in vocal feature extraction;

(3) Utilized Wave and Librosa packages to acquire sound amplitude and frequency data, respectively, computing loudness variability and pitch values using the formula proposed by Fu *et al.* (2024).

$$Loudness\_varibililty = \sqrt{\frac{\sum_{n\in N}(SA_n - Mean(SA_N)}{|N|}} \quad (1)$$

where n represents audio frames per second;
N represents the set of audio frames for 2-second slices;
SA represents the sound amplitude of the audio frame;
|N| is the number of audio frames.

$$Vocal\_pitch = \frac{\sum_{n\in N} Frequency(b)}{|N|} \quad (2)$$

where b represents a particular frequency band;
Frequency(b) denotes the fundamental frequency of b (F0);

For background audio in travel vlogs, this study extracted Spectral Centroid (SpeCentroid) by leveraging the librosa library. Python was employed to invoke librosa.feature.spectral_centroid() for acquiring time-series data and sampling rates per audio frame, thereby calculating spectral centroids for individual frequencies. For 2-second video segments, the SpeCentroid value was determined by averaging all frame-level centroids within the segment. Similarly, librosa.feature.zero_crossing_rate() was executed via NumPy to compute Zero-Crossing Rate (ZCR) by quantifying zero-crossing events per audio frame.

For Danmaku annotations, this study imported Python crawled data into a Danmaku-specific database. The number of Danmaku comments was calculated by quantifying the number of occurrences of the string using the built-in aggregation function COUNT( ). Additionally, to examine whether facial presence moderates audio features' impact on audience participation, we introduced a binary moderator variable face: coded as 1 if human faces appeared in the 2-second video segment and 0 otherwise. Variable definitions and descriptive statistics are presented in Table 2.

Table 2：Descriptive statistics

Variant	Min	Max	Mean	Std
Face	.000	1.000	.345	.476
Pitch	129.549	578.513	323.730	64.337
Loudness	.009	.220	.077	.039
Spe Centroid	359.664	4928.560	2048.666	565.604
ZCR	.016	.409	.091	.038

Empirical Model

We specified the baseline model shown in Eq. (1) as model 1 to test the main hypotheses. And we added the interaction terms into Eq. (1) as model 2 to test the moderating effect, which is shown in Eq. (2). All variables were standardized with the Z-score standardization method.

Number of Danmaku = $\beta_0 + \beta_1 Loudness + \beta_2 Pitch + \beta_3 SpeCentroid + \beta_4 ZCR$ (1)

Number of Danmaku = $\beta_0 + \beta_1 Loudness + \beta_2 Pitch + \beta_3 SpeCentroid + \beta_4 ZCR + \beta_2 Loudness \times Face + \beta_2 Pitch \times Face + \beta_3 SpeCentroid \times Face + \beta_4 ZCR \times Face$ (2)

RESULTS

Before conducting hypothesis testing, we performed a multicollinearity test on the model. In regression analysis, when there is a high correlation between the explanatory variables, it may result in multicollinearity issues, which affect the stability of the model. Generally, the tolerance and the variance inflation factor (VIF) are used as diagnostic indicators for collinearity. When the tolerance value is greater than 0.1, the mean VIF is less than 2, and the maximum VIF is less than 10, it indicates that no collinearity problem exists. We tested the tolerance value and the VIF coefficients of the research model, and the results show that the tolerance values are all greater than 0.78, meeting the minimum requirement. Furthermore, the VIFs of variables range from 1.017 to 1.267, all of which are less than 10, with an average of 1.132, suggesting no multicollinearity.

Hypothesis Testing

The results of the logistic regression are shown in Table 3. Column (1) shows that both human-acoustic and background sound cues have a significant effect on user participation. For human-acoustic cues, the loudness of the human voice is not significantly associated with user participation. Thus, H1a is not supported. The pitch of the human voice is positively related to the user participation with a coefficient of 0.116 ($p < 0.001$). Thus, H1b is supported. In terms of background vocal cues, SpeCentroid was positively correlated with user participation with a coefficient of 0.241 ($p < 0.001$), and thus H2a is supported. However, ZCR had a negative impact on user participant with a coefficient of -0.197 ($p < 0.001$). Therefore, H2b is supported. However, ZCR had a negative impact on user participation with a coefficient of -0.197 ($p < 0.001$). Therefore, H2b is not supported. Column (2) shows that the effect of loudness is negatively moderated by the presence or absence of faces, with an interaction term coefficient of -0.188. Meanwhile, faces positively affect the relationship between ZCR and user participation, with an interaction term coefficient of 0.230. The interaction effect between pitch and faces is positive but not statistically significant. Similarly, the interaction effect of SpeCentroid with faces was negative but not statistically significant.

Table 3：Regression results

	Dependent Variable: Number of Danmaku	
	Model 1	Model2
Loudness	-0.012	0.069*
Pitch	0.116***	0.096**
SpeCentroid	0.241***	0.309***
ZCR	-0.197***	-0.320***
Face		-0.063
Loudness × Face		-0.188***
Pitch × Face		0.045
Spe Centroid × Face		-0.135
ZCR × Face		0.230**
Note: *p < 0.05, **p < 0.01, ***p < 0.001		

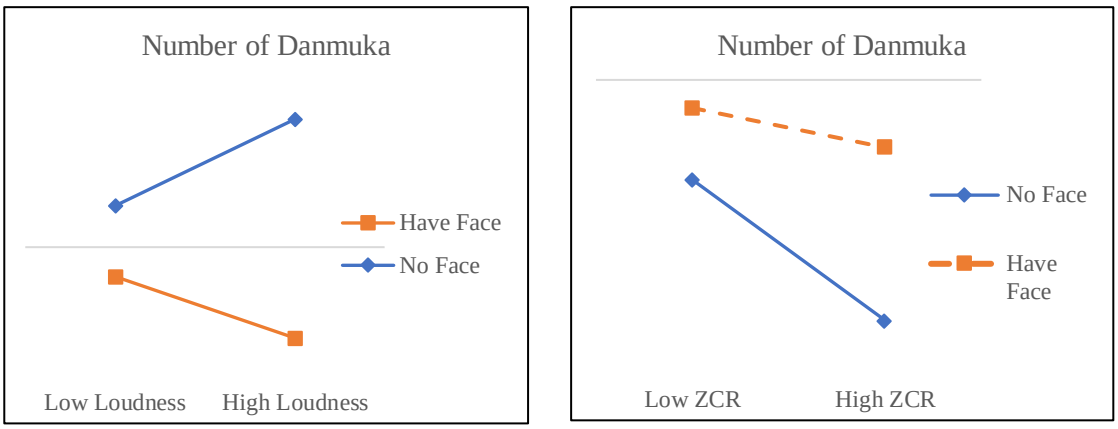

Figure 3： Moderating effect of face on the relationship between acoustic cues and user participation

DISCUSSION AND CONCLUSION

Discussion

This study focuses on factors influencing audience behavioral participation in travel vlogs within short-video platforms. Through comprehensive literature review, we identified acoustic features and facial presence as critical determinants of participation. Acoustic features were categorized into vocal and background audio features, with facial presence serving as a moderator variable to examine how these factors collectively impact behavioral participation.

Regarding acoustic effects on travel vlog audience participation without facial presence, vocal Pitch and background SpeCentroid positively enhance participation. However, ZCR shows a significant negative effect (contradicting H2b). As ZCR measures zero-crossing frequency per time unit—positively correlated with auditory roughness and acoustic complexity—low ZCR yields smooth audio, while high ZCR produces harsh/chaotic sounds. Elevated background ZCR generates high-frequency noise, inducing auditory fatigue and attentional diversion or masking vocal content. These effects hinder content focus, ultimately reducing participation. Furthermore, H1a was refuted, indicating that vocal pitch characteristics ("how" content is delivered) exert a stronger influence on audience participation than loudness amplitude ("how loudly"). This suggests pitch effectively conveys paralinguistic dimensions of emotion and informational salience , enhancing viewer resonance. In contrast, loudness exhibits semantic ambiguity: high amplitude may signal enthusiasm/excitement or perceived noisiness/aggression; low amplitude may indicate gentleness/composure or apathy/passivity. This polysemy introduces individual-level variance in emotional decoding, ultimately diluting loudness's statistical significance in participation models.

The introduction of face significantly altered the pathway through which acoustic cues influence user participation. Regarding loudness-participation relationships, no significant association was observed without a face, while with a face, the effect direction reversed to a significant negative correlation. This indicates face visibility reshapes users' perceptual logic toward vocal loudness—excessive loudness may be interpreted as emotional hyperactivity or strong interference when faces are present, reducing interaction intention. Similarly, face moderates ZCR-participation associations: the negative effect weakens with face but intensifies without face. The underlying mechanism suggests faces, as visual symbols of social presence, alleviate negative experiences from chaotic background sounds (high ZCR) through perceived co-presence, enhancing tolerance for acoustic instability. Conversely, absent face, such auditory chaos is attributed to poor content quality, significantly suppressing interaction.

Theoretical Implications

This study transcends traditional unimodal frameworks by incorporating face as a moderator alongside acoustic features, revealing cross-modal interactions that drive audience participation in travel short videos. Acoustic cues (vocal/background audio) and visual cues (face) dynamically interact to shape participation decisions, demonstrating that sound influences behavior not in isolation but through multimodal integration. This advances empirical evidence for cross-modal moderation in multimedia communication while pioneering exploration of other modality combinations (e.g., visual composition/dynamic effects with acoustics). Simultaneously, it refines acoustic boundaries by delineating differential impacts within vocal dimensions (Pitch vs. Loudness) and background audio (SpeCentroid vs. ZCR), confirming acoustic heterogeneity in behavioral influence and advancing granular feature-specific analysis for the sensory stimuli-behavioral response pathway in communication psychology.

Practical Implications

Tourism short video creators should optimize audio-visual synergy by prioritizing Pitch expressiveness over Loudness amplitude refinement for vocal delivery and strictly controlling vocal loudness during face exposure to prevent negative perceptions. For background audio, select high-SpeCentroid tracks to boost participation while suppressing ZCR in face-absent content to avoid participation reduction from high-frequency switching sound effects. Strategically calibrate face exposure: maximize visibility for emotional resonance/social interaction-focused content to enhance cross-modal synergy, but balanced exposure for immersive experience/information delivery-oriented content to mitigate acoustic interference through scenario-specific planning.

Limitations and Future Research

This study deliberately confined its focus to the core triad of vocal-background audio-face interactions, excluding richer acoustic features (e.g., rhythmic patterns, timbre) and visual elements (e.g., color grading, dynamic effects), leaving substantial scope for expanding cross-modal exploration. Future research should integrate physiological metrics to decode users' cognitive processing of multimodal stimuli and extend investigations to diverse communication contexts (e.g., long-form videos, game live-streaming) to validate the generalizability of findings, thereby advancing theoretical frameworks for user interaction in multimedia dissemination.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (72361034) and the Humanities and Social Science Foundation of Yunnan University (2023YNUGSP06, 2022YNUGSP05).

REFERENCES

Aeschlimann, M., Knebel, J. F., Murray, M. M., & Clarke, S. (2008). Emotional pre-eminence of human vocalizations. *Brain Topography, 20*, 239-248. https://doi.org/10.1007/s10548-008-0051-8

Bai, Q., Hu, Q. V., Ge, L., & He, L. (2019). Stories that big Danmaku data can tell as a new media. *IEEE Access, 7*, 53509-53519. https://doi.org/10.1109/ACCESS.2019.2909054

Bai, Q., Wei, K., Zhou, J., Xiong, C., Wu, Y., Lin, X., & He, L. (2021). Entity-level sentiment prediction in Danmaku video interaction. *The Journal of Supercomputing, 77*(9),9474-9493. https://doi.org/10.1007/s11227-021-03652-4

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision, 9*(12),10-10. https://doi.org/10.1167/9.12.10

Charest, I., Pernet, C. R., Rousselet, G. A., Quiñones, I., Latinus, M., Fillion-Bilodeau, S., *et al.* (2009). Electrophysiological evidence for an early processing of human voices. *BMC neuroscience, 10*(1), 127. https://doi.org/10.1186/1471-2202-10-127

Dai, X., & Wang, J. (2024). The effect of video highlights on the intention to give free virtual gifts. *Electronic Commerce Research and Applications, 63*, 101342. https://doi.org/10.1016/j.elerap.2023.101342

Flores-Saviaga, C., Hammer, J., Flores, J. P., Seering, J., Reeves, S., & Savage, S. (2019). Audience and streamer participation at scale on Twitch. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, 277-278. https://doi.org/10.1145/3342220.3344926

Fu, S., Wu, Y., Du, Q., Li, C., & Fan, W. (2024). The secret of voice: How acoustic characteristics affect video creators' performance on Bilibili. *Decision Support Systems, 179*, 114167. https://doi.org/10.1016/j.dss.2023.114167

Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology, 5*(1), 1-29. https://doi.org/10.1207/s15326969eco0501_1

Gray, R. M., & Markel, J. D. (1976). Distance measures for speech processing. IEEE Transactions on *Acoustics, Speech, and Signal Processing, 24*(5), 380-391. https://doi.org/10.1109/tassp.1976.1162849

Guido, G., Pichierri, M., Pino, G., & Nataraajan, R. (2019). Effects of face images and pareidolia on consumers' responses to print advertising: An empirical investigation. *Journal of Advertising Research, 59*(2), 219-231. https://doi.org/10.2501/JAR-2018-030

Hazmoune, S., & Bougamouza, F. (2024). Using transformers for multimodal emotion recognition: Taxonomies and state of the art review. *Engineering Applications of Artificial Intelligence, 133*, 108339. https://doi.org/10.1016/j.engappai.2024.108339

He, Y., & Tang, T. Y. (2017, September). Recommending highlights in Anime movies: Mining the real-time user comments "Danmaku". *2017 Intelligent Systems Conference (IntelliSys)*, 319-322. https://doi.org/10.1109/intellisys.2017.8324311

Li, J., Wang, Y., Lin, Y., Jiang, Y., & Yang, Q. (2025). The power of sound: The impact of auditory features in hotel short influencer-generated videos on viewer engagement. *International Journal of Hospitality Management, 131*, 104341. https://doi.org/10.1016/j.ijhm.2025.104341

Lin, X., Li, X., & Mou, J. (2024). Exploring user engagement behavior with short-form video advertising on short-form video platforms: A visual-audio perspective. *Internet Research*. https://doi.org/10.1108/INTR-07-2023-0521

Luo, L., Liu, L., Zheng, Y., & Wang, Y. (2025). The power of voice: Investigating the effects of streamer voice characteristics on sales performance in live streaming E-commerce. *Journal of Retailing and Consumer Services, 84*, 104260. https://doi.org/10.1016/j.jretconser.2025.104260

Ni W, & Coupe C. (2023). Time-synchronic comments on video streaming website reveal core structures of audience engagement in movie viewing. *Frontiers in Psychology, 13*, 1040755. https://doi.org/10.3389/fpsyg.2022.1040755

Přibil, J., & Přibilová, A. (2011). Statistical analysis of complementary spectral features of emotional speech in Czech and Slovak. *Text, Speech and Dialogue*, 6836. https://doi.org/10.1007/978-3-642-23538-2_38

Rutten, S., Santoro, R., Hervais-Adelman, A., Formisano, E., & Golestani, N. (2019). Cortical encoding of speech enhances task-relevant acoustic information. *Nature Human Behaviour, 3*, 974-987. https://doi.org/10.1038/s41562-019-0648-9

Shi, Y. H., Zhao, Y. F., & Zhao, C. Y. (2024). The impact of Danmaku comments on the sense of virtual community. *Psychological Techniques and Application*s, 12(12), 713-722. https://doi.org/10.16842/j.cnki.issn2095-5588.2024.12.002 (in Chinese).

Shilaskar, S., Bhatlawande, S., Shinde, S., & Sattigeri, S. (2023). That classification using recurrent neural networks with long

short-term memory and support vector machine. *International Journal on Recent and Innovation Trends in Computing and Communication, 11*(10), 1377-1388. https://doi.org/10.17762/ijritcc.v11i10.8680

Simmonds, L., Bogomolova, S., Kennedy, R., Nenycz-Thiel, M., & Bellman, S. (2020). A dual-process model of how incorporating audio-visual sensory cues in video advertising promotes active attention. *Psychology & Marketing, 37*(8), 1057-1067. https://doi.org/10.1002/mar.21357

Sun, S., Wang, F., & He, L. (2018). Movie summarization using bullet screen comments. *Multimedia Tools and Applications, 77*, 9093-9110. https://doi.org/10.1007/s11042-017-4807-6

Takeuchi, A., & Nagao, K. (1993). Communicative facial displays as a new conversational modality. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*, 187-193. https://doi.org/10.1145/169059.169156

Wang, C., & Li, Z. (2024). Impact of discrete emotions on audience engagement with climate change videos on Chinese TikTok (Douyin). *Social Behavior and Personality: An International Journal*, *52*(4), 1-12. https://doi.org/10.2224/sbp.13076

Wood, M. K. (2019). What makes a vlog a vlog. *Diggit Magazine.* Retrieved from https://www.diggitmagazine.com/academic-papers/what-makes-vlog-vlog (accessed 18 May 2025).

Xu, D., Chen, T., Pearce, J., Mohammadi, Z., & Pearce, P. L. (2021). Reaching audiences through travel vlogs: The perspective of involvement. *Tourism Management, 86*, 104326. https://doi.org/10.1016/j.tourman.2021.104326