

Vocalize: Lead Acquisition and User Engagement through Gamified Voice Competitions

Edvin Teskeredzic
Infobip
Sarajevo, Bosnia
Edvin.Teskeredzic@infobip.com

Muamer Paric
Infobip
Sarajevo, Bosnia
Muamer.Paric@infobip.com

Adna Sestic
Infobip
Sarajevo, Bosnia
Adna.Sestic@infobip.com

Petra Fribert
Infobip
Zagreb, Croatia
Petra.Fribert@infobip.com

Anamarija Lukac
Infobip
Zagreb, Croatia
Anamarija.Lukac@infobip.com

Hadzem Hadzic
Infobip
Sarajevo, Bosnia
hadzem.hadzic@infobip.com

Kemal Altlwkany
Infobip
Sarajevo, Bosnia
kemal.altlwkany@infobip.com

Emanuel Lacic
Infobip
Zagreb, Croatia
emanuel.lacic@infobip.com

Abstract

This paper explores the prospect of creating engaging user experiences and collecting leads through an interactive and gamified platform. We introduce Vocalize, an end-to-end system for increasing user engagement and lead acquisition through gamified voice competitions. Using audio processing techniques and LLMs, we create engaging and interactive experiences that have the potential to reach a wide audience, foster brand recognition, and increase customer loyalty. We describe the system from a technical standpoint and report results from launching Vocalize at 4 different live events. Our user study shows that Vocalize is capable of generating significant user engagement, which shows potential for gamified audio campaigns in marketing and similar verticals.

CCS Concepts

• **Human-centered computing** → **User models**; *Natural language interfaces*; • **Computing methodologies** → Speech recognition.

Keywords

lead acquisition, user engagement, gamification, generative AI, speech processing

ACM Reference Format:

Edvin Teskeredzic, Muamer Paric, Adna Sestic, Petra Fribert, Anamarija Lukac, Hadzem Hadzic, Kemal Altlwkany, and Emanuel Lacic. 2025. Vocalize: Lead Acquisition and User Engagement through Gamified Voice Competitions. In *Adjunct Proceedings of the 36th ACM Conference on Hypertext and Social Media (HT Adjunct 2025)*, September 15–19, 2025, Chicago, IL, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3720533.3750059>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

HT Adjunct 2025, Chicago, IL, USA

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1533-4/25/09

<https://doi.org/10.1145/3720533.3750059>



Figure 1: Example of a Vocalize campaign launched at WeAre-Developers 2024.

1 Introduction

The active interaction of users with digital content has become a central pillar of success in modern business [1]. Active user engagement results in increased customer loyalty for existing clients [15], as well as the acquisition of new leads, often of high quality. However, generating user engagement through experiences that are both engaging and organic can be a challenging task for many organizations. One of the main hurdles in creating engagement, both for existing and new customers, is the lack of motivation and personalization when interacting with content [7]. A possible solution to these problems comes in the form of user participation through gamified competitions.

Voice-driven chatbot interfaces, when combined with gamification concepts, can transform how businesses acquire leads and improve brand awareness. Such an approach makes interactions more engaging, and ensures personalized experiences that effectively capture attention and build lasting customer relationships [2]. As demonstrated in prior research, gamified conversational agents have shown potential to boost motivation in educational contexts [4],

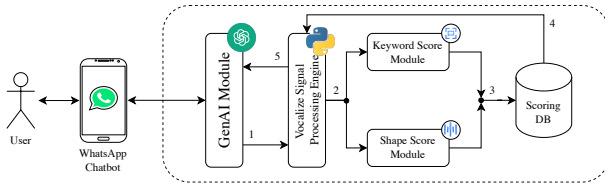


Figure 2: The underlying components of Vocalize. User audio inputs via WhatsApp are processed through keyword and shape scoring modules, with results enhanced by a generative AI and stored in a scoring database for feedback and leaderboard updates.

while voice-based agents have proven effective in crafting immersive narratives [14]. Furthermore, frameworks leveraging dynamic language adaptation and conversational cues have been shown to foster emotional engagement and trust, maintaining user-centered interactions [3, 10].

Contribution. In this work we present *Vocalize*¹, an end-to-end system for increasing user engagement and lead acquisition through gamified voice competitions. Vocalize offers a multimodal interface that allows users to communicate with its underlying chatbot using their voice and text. By combining audio-based gamification and generative AI-based content personalization, we show how to improve lead acquisition and increase customer engagement, which in turn can result in greater brand awareness [13]. We support our claims by analyzing the results obtained in lead acquisition and user engagement from 4 venues at which we launched Vocalize-supported campaigns.

2 Vocalize

For simplicity, we will refer to end users who interact with Vocalize as *users*, while the term *brand* is used for those designing and launching specific Vocalize campaigns. The main purpose of Vocalize is simple; enable brands to acquire leads of higher quality and have their users more engaged by providing them a gamified experience, with brands optionally handing out prizes to best ranked users.

Vocalize is built on top of WhatsApp² as its underlying channel for communication. WhatsApp has been chosen as it is the most popular messaging application in terms of monthly active users - more than 3 billion [12]. To do this, we utilize Infobip’s Answers platform³ for integrating with WhatsApp as well as to access OpenAI’s commercially available LLMs [8] to provide generative AI responses.

Competition outline. To initialize a competition, brands are only required to come up with a textual catch phrase and a desired image. For example, for a panoramic image of a city it could be the city skyline, or for an object it could be the edges of the object. Once users start interacting with the conversational agent over WhatsApp, they are provided with the predefined catch phrase and

the outline of the image. Their task is then to record an audio message in which they repeat the catch phrase, but with an additional twist: the shape of the recorded audio message must also match the outline as close as possible. An example can be seen in Figure 1 where the catch phrase "I love Berlin" was used in addition to the image that represents the skyline of Berlin. Here we further refer to the outline of that image as contour. Figure 1 also shows how a user can improve their rank by recording an audio message that better matches the given contour. In summary, the main challenge of Vocalize in which users compete is to record an audio message that matches the campaign catch phrase in terms of content (spoken words), while also trying to match the shape of their audio recording (as visible in WhatsApp) with the target contour.

System architecture. A complete overview of Vocalize is provided by Figure 2. The WhatsApp chatbot interface enables users to communicate with Vocalize using natural language. If a user decides to participate in Vocalize’s main challenge and sends an audio recording, it is forwarded to Vocalize’s internal signal processing engine. The signal processing engine uses two separate components to formulate the score for the given user’s audio recording: a keyword scoring module and a shape scoring module. The user’s score is then logged to the database and the user is provided with feedback information. An example of such an interaction is shown in Figure 3.

2.1 Chatbot Interface

A typical way brands promote Vocalize is by advertising a QR code that initiates a conversation with Vocalize. This is useful for both virtual promotions, but also live venues as brands can print out the QR code on booths, posters or merchandise.

Once an interaction has been initiated, the conversational agent explains the rules of the game and collects contact information. As seen in Figure 3, the generative AI model enables the user to interact

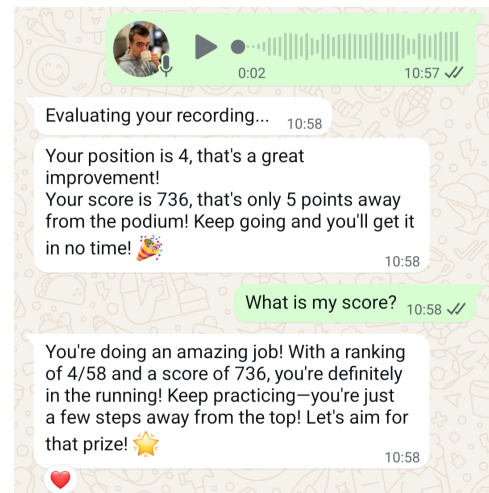


Figure 3: Users can interact with the system using natural language to receive scores, feedback, and personalized guidance in real time.

¹<https://www.infobip.com/ai-hub/vocalize>

²<https://www.whatsapp.com/>

³<https://www.infobip.com/docs/answers>

Table 1: Statistics of the live events in 2024 where the user study was conducted. Larger events like WeAreDevelopers and Web Summit attracted significantly more voice recordings and longer interaction durations, while all events mostly demonstrated consistent message durations and engagement formats.

Event	Date	Target image	Target phrase	Voice recordings	Total duration	Median duration
WeAreDevelopers	17.07. - 19.07.	Skyline of Berlin	I love Berlin	6,321	4h 41m	2.25s
KulenDayz	30.08. - 01.09.	Skyline of Osijek	I love Kulen	1,257	46m	2.05s
GOTO Chicago	21.10. - 22.10.	Skyline of Chicago	Go to Infobip	1,216	1h 42m	3.38s
Web Summit	11.11. - 14.11.	Skyline of Lisbon	I love Lisbon	3,662	2h 22m	2.69s

with the system using natural language. For example, the user can ask the chatbot about their score or about the prizes without having to click through menus. If the user sends an audio message, it is automatically forwarded to Vocalize’s underlying signal processing engine which assigns each recording a score. The score is returned, along with relevant metadata, e.g.: user’s position on leaderboard, number of attempts made, next best score, etc. The generative AI module is used to polish the metadata and present it to the user in a more natural form that mimics a real conversation.

2.2 Signal Processing Engine

Vocalize relies on an underlying signal processing engine to assign scores to user recordings. It consists of two components, both of which produce an individual score for any given audio: keyword score and shape score. A Vocalize campaign can have one or both components active for scoring, which is defined by the brand.

Keyword scoring module. This module rates how well the user’s spoken content matches the reference catch phrase (i.e., a target phrase). The module outputs a score ranging from 0 (total mismatch) to 1 (perfect match). To determine what the user said, we obtain a transcript of the user’s audio recording via automatic speech recognition. Depending on the campaign and brand requirements such as legal or privacy related considerations, we utilize an open-source solution, such as Whisper [9] or opt for a proprietary speech-to-text provider. The score itself is then computed by slightly modifying the Levenshtein algorithm for computing the distance between two strings [5]:

$$S = 1 - \frac{D}{L}, \quad (1)$$

where D is the Levenshtein distance between the two strings and L is the length of the longer of the two strings (target phrase and user transcript). If the two strings do not share characters in common, this score will equal 0, while if they are identical, the score will be 1, which directly matches our requirements.

Shape scoring module. This module computes the visual similarity between the shape of the user audio recording (more precisely the envelope of its waveform) and the campaign contour. Users only have visual access to the audio waveform as shown in WhatsApp’s interface: segmented into 40 vertical bars. Thus, we adjusted the shape scoring module to accommodate for this. First, we segment both the contour and the audio waveform of the user temporally into 40 equally spaced segments. We compute the root mean square of each of these segments, thereby producing two temporal vectors of equal length (40) with non-negative values. Computing the similarity between these two vectors is then done using the dot product which, given the non-negativity of the vectors’ elements, results in

a value between 0 and 1, thereby matching our requirements and being in scale with the keyword scoring module.

Brands may choose which algorithm to use. We have not performed a detailed analysis of the performance differences between the two approaches and plan to do so in future work. Our speculation is that the first approach more closely resembles what humans consider to be “similar”.

2.3 Generative AI

The use of the generative AI module as part of Vocalize is twofold:

Enhanced UX. By leveraging intent detection, we enable the user to communicate with Vocalize using natural language. The user can query the system about updates regarding the competition (i.e. when does the competition close, what the user’s current ranking is, etc.). We enable such communication by evaluating the embeddings of user messages, and classifying those into pre-determined intents. These intents are then mapped to Vocalize API calls, with the API response being fed to an LLM (in our case, OpenAI ChatGPT [8] family of models) in order to transform the JSON data into a natural language response, which gets presented to the user.

User engagement. Vocalize uses a predefined conversational agent adapted through a system prompt ⁴, when informing the user about their score after their attempt (i.e., after sending a voice message). The assistant is instructed to keep the user engaged, by offering encouragement, giving hints and tips, and using the current user’s stats (i.e., number of attempts, best score, points to reach podium, etc.). Our goal is to increase user engagement while maintaining a fun, light-hearted tone when communicating with the participants.

3 Demo and User Study

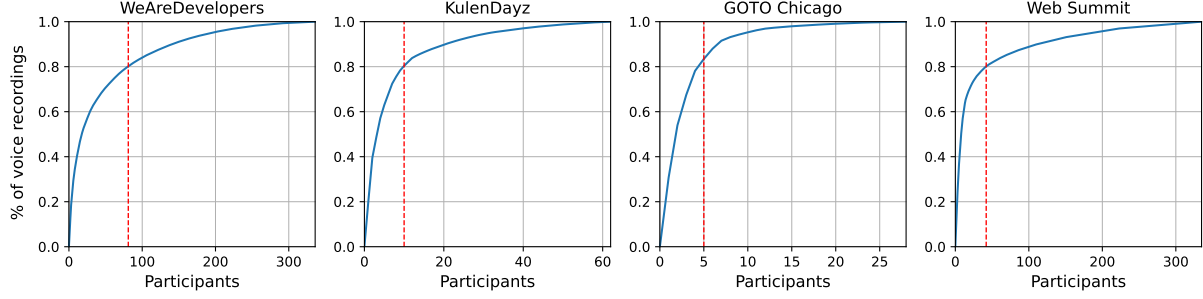
In order to validate the efficacy of Vocalize, we conducted a user study across two larger (i.e., WeAreDevelopers and Web Summit Lisbon) and two smaller (i.e., Kulendayz and GOTO Chicago) live events in 2024. Here, our aim was to demonstrate the benefits of using gamified audio competitions in terms of user engagement, lead acquisition, and user experience.

As seen in Table 1, each competition incorporated unique target phrases aligned with the event theme as well as the skyline of the respective city of the event as the target image (also seen in Figure 1). The duration of the audio messages between all events ranged from 0.5 to 16.8 seconds. In addition, larger events did attract more participation with respect to the number of voice recordings and the total time of all voice messages recorded. The overall performance

⁴The prompt message contained instructions to respond in an engaging, fun and gamified way and to provide the participant with instructions on how to compete and possibly win in the competition.

Table 2: Lead acquisition progression for all possible participants that sent an initial message (i.e., *Potential leads*) to the chatbot.

	Potential leads	Leads	Participants	Recurring participants	Textual messages	Audio messages
WeAreDevelopers	430	71.16%	68.60%	64.42%	25%	75%
KulenDayz	66	71.21%	69.69%	69.69%	21%	79%
GOTO Chicago	35	71.43%	65.71%	60.00%	13%	87%
Web Summit	794	59.57%	42.32%	27.58%	65%	35%

**Figure 4: Cumulative frequency of voice recordings. Roughly 80% of all voice messages (red) were sent by 24.11% of participants in case of WeAreDevelopers, 16.13% in case of KulenDayz, 17.86% in case of GOTO Chicago and 12.54% of participants at Web Summit.**

of each voice-based competition is reported in Table 2. The reported proportion of audio messages reflect the number of voice recordings shown in Table 1. Overall, we see a high participation rate once a user has sent an initial message (e.g., a *potential lead* becomes a *lead* once the user registration for the competition is finished, and a *participant* if at least one voice recording has been sent). The progression from *potential leads* to *recurring participants* highlights the differences in user engagement across events. The smaller events demonstrated strong transition probabilities. In contrast, the larger events exhibited more varied outcomes. While WeAreDevelopers maintained high levels of recurring participation, Web Summit showed a notable drop-off. That is, at Web Summit we see a much higher proportion of textual messages when compared to voice recordings. We hypothesize that the reason for such a behavior lies in the chatbot flow that was additionally extended for Web Summit - a Retrieval-Augmented Generation (RAG) component [6] that invited participants of the conference (i.e., without the need of completing a registration) to learn more about Vocalize. Although, still fulfilling the purpose of brand awareness, such an addition may have inadvertently directed participants away from the competition and reduced the prominence of audio-based engagement.

Participation Dynamics. With respect to individual user engagement, as shown in Figure 4, a small subset of highly active participants contributes the most to the total volume of recorded voice messages. This almost matches with the Pareto Principle [11] as the distribution of recorded voice messages does have characteristics of a power-law distribution. For example, the most active participant at WeAreDevelopers recorded 460 voice messages, at KulenDayz it was 259, for GOTO Chicago 381 and 298 at Web Summit. Our findings suggest that voice-based competitions in combination with conversational agents can resonate strongly with customers at live events.

4 Conclusion and Future Work

In this paper, we presented Vocalize - an end-to-end system for building voice-enabled competitions, based around the concept of matching a voice recording to a given target phrase and image. Furthermore, we gave a detailed breakdown of the main components of Vocalize from a technical standpoint. To assess the impact on user engagement and lead acquisition, we further evaluated our method across four live events where we demonstrated a strong participation rate and conversion from potential leads to recurring participants.

We plan to enable the use of Vocalize with multiple communication channels, as well as customization of campaigns for use in marketing, event management, and exhibitions. Here we plan to also investigate how the introduction of other novel audio scoring mechanisms in voice-based competitions impacts user engagement. In addition, we intend to explore whether age, cultural, and industry background play an important role in engaging with gamified user experience by conducting further user studies at events aimed for different backgrounds.

References

- [1] Surajit Bag, Gautam Srivastava, Md Mamoon Al Bashir, Sushma Kumari, Mihalios Giannakis, and Abdul Hannan Chowdhury. 2022. Journey of customers in this digital era: Understanding the role of artificial intelligence technologies in user engagement and conversion. *Benchmarking: An International Journal* 29, 7 (2022), 2074–2098.
- [2] Abid Haleem, Mohd Javaid, Mohd Asim Qadri, Ravi Pratap Singh, and Rajiv Suman. 2022. Artificial intelligence (AI) applications for marketing: A literature-based study. *International Journal of Intelligent Networks* 3 (2022), 119–132.
- [3] Rutuja Joshi and Klaus Bengler. 2024. Crafting Human-AI Interaction: A Rhetorical Approach to Adaptive Interaction in Conversational Agents. In *Proceedings of the 12th International Conference on Human-Agent Interaction*. 314–322.
- [4] Bijan Khosrawi-Rad, Arne Borchers, Linda Grogoric, and Susanne Robra-Bissantz. 2023. Design Principles for Gamified Pedagogical Conversational Agents. In *Proceedings of the 26th International Academic Mindtrek Conference*. 67–82.
- [5] Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. Soviet Union, 707–710.

- [6] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
- [7] Heather L O'Brien and Elaine G Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology* 59, 6 (2008), 938–955.
- [8] R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. View in Article 2 (2023), 3.
- [9] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.
- [10] Jude Rayan, Dhruv Kanetkar, Yifan Gong, Yuewen Yang, Srishti Palani, Haijun Xia, and Steven P Dow. 2024. Exploring the potential for generative ai-based conversational cues for real-time collaborative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*. 117–131.
- [11] Robert Sanders. 1987. The Pareto principle: its use and abuse. *Journal of Services Marketing* 1, 2 (1987), 37–40.
- [12] Sinch. 2025. Most Popular Messaging Apps by Country. <https://sinch.com/blog/most-popular-messaging-apps-by-country/> Accessed: 2025-06-01.
- [13] Sang Vo Minh, Giang Nguyen Huong, and Giang Dang Nguyen Ha. 2022. The role of social brand engagement on brand equity and purchase intention for fashion brands. *Cogent Business & Management* 9, 1 (2022), 2143308.
- [14] Ying Xu, Valery Vigil, Andres S Bustamante, and Mark Warschauer. 2022. “Elinor’s talking to me!”: Integrating conversational AI into children’s narrative science programming. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–16.
- [15] Xiabing Zheng, Christy MK Cheung, Matthew KO Lee, and Liang Liang. 2015. Building brand loyalty through user engagement in online brand communities in social networking sites. *Information Technology & People* 28, 1 (2015), 90–106.