Moderation challenges in digital gaming spaces: Prevalence of offensive behaviors in voice chat

Rachel Kowert

Take This

Liz Woodwell

Modulate

Executive Summary

- Content moderation is an ongoing challenge in digital gaming spaces, particularly the moderation of voice chat content.
- To better understand the landscape of the verbal space within games, Take This partnered with Modulate to develop the first known baseline of information about the nature of offensive behaviors within the voice-chat of gaming spaces.
- Analyses revealed that 1 in 4 (26.43%) of all players had at least one incidence of offensive language.
- Among all players, 21.39% were flagged with only non-severe offenses and 5.03% were flagged with at least one severe offense over the last 30 days.
- Racial/cultural hate speech was the most common offense, constituting more than half of all offenses by all users by category.
- Racial/cultural and gender/sexual hate speech were more likely to be non-severe than severe offenses, suggesting a normalization of these specific kinds of hate speech within gaming cultures.
- Sexual vulgarity was the only type of offense that was more likely to be severe than non-severe, suggesting a potentially accelerated trajectory to this kind of behavior.
- Perceived adult players were more likely to be flagged with an incidence of offensive speech (36%) than perceived underage players (17%).
- Perceived adult players are more than twice as likely than perceived underage players to have at least one severe offense.

Content moderation in gaming spaces: Current challenges and trends

Content moderation is an ongoing challenge in digital gaming spaces. Broadly speaking, content moderation refers to the efforts taken to ensure that user-generated content (e.g., chat) within any particular environment adheres to the space's rules, guidelines, and terms of service. The ways in which the content is moderated can vary, from fully automated or AI systems to strategies that rely on human moderation, and everything in between. The particular moderation guidelines followed by any company or platform can vary. As such, so does the effectiveness of different moderation approaches.

Digital games and gaming platforms have historically struggled with implementing effective moderation strategies for the text-based, user generated chat within them. This is evidenced by the high prevalence rates of hate, harassment, and other offensive behaviors that are well-documented within these environments. A 2022 report from the Anti-Defamation League (ADL) indicated that 83% of adults and 60% of teenagers experienced harassment in online multiplayer games in the last six months. Of these, 71% of adult online multiplayer game players experienced severe abuse, including physical threats, stalking, and sustained harassment. Similarly, Kowert and Cook (2022) reported that nearly half of all game players have directly experienced sexual harassment (45%) and violent threats (46.8%) in-game, and more than half have experienced hate speech (64%). The prevalence of these experiences have led some scholars to argue that hate speech has come to be a culturally justified behavior within games and gaming cultures under the umbrella of "toxic gamer cultures" (Consalvo, 2012; Paul, 2018).

As it stands today, the most common moderation strategies within gaming spaces are text-based moderation. This type of moderation focuses on the monitoring of live text-based chat sessions through human and/or automated detection of keywords or phrases to determine whether users are engaging in inappropriate behavior as set by rules and guidelines of any particular gaming space. The effectiveness of this kind of moderation varies. While it can be effective at flagging the most objectionable content and obvious offenses, this kind of text-based language moderation will inevitably always be one step behind as language and vocabulary adapt to thwart moderation attempts. Furthermore, text-based moderation strategies do not include the analysis of voice (i.e., verbal) communication between users.

Voice chat within gaming spaces (sometimes referred to as VoIP, voice over internet protocol) has been a cornerstone of gaming spaces for the last two decades. Today, most digital game systems and platforms have integrated voice chat as a standard feature (e.g., Xbox, Playstation) and many players use third party services (e.g., Discord) for voice-chat functionality while playing digital games. While it is difficult to find specific statistics around how many users utilize voice chat in gaming spaces, it is believed to be nearly ubiquitous (Williams, 2015; Ross, 2021; Subspace, 2021). An industry report from Tencent noted 90.6% of their players use the built-in voice chat function when playing and, when a title doesn't have an in-game voice communication system in place, 73.7% of players said they utilize a third-party service for this functionality (Tencent, 2022).

Despite the fact that most game players seem to be utilizing voice chat to communicate within gaming spaces, it has received little attention when it comes to moderation efforts. As a consequence, it remains unclear to what extent verbal communication reflects the same levels of hate and harassment as has been documented in text-based communication within gaming spaces.

Landscape of voice chat in games

To better understand the landscape of the verbal space within games, Take This partnered with Modulate to develop the first known baseline of information about the nature of offensive behaviors within the voice-chat of gaming spaces. The primary goal of this collaboration was to evaluate the prevalence of offensive voice chat content within game spaces to better understand the landscape of verbal exchanges within gaming environments.

Modulate is a Boston-based company with the mission is to build a richer, safer and more inclusive online space through "proactive voice moderation". Their tool, called ToxMod, provides real-time analysis of voice chat to allow content moderators to swiftly take appropriate action within the largely unmoderated space of verbal communication. ToxMod is unique as it not only understands the words that are being said but also analyzes context in order to determine whether what is being said is being done with intent to harm. This is critical, as the context of language is particularly important within gaming spaces. For example, in a video game "I am going to kill you" is something that is commonly said in nearly any competitive game. As such, a text-based, keyword moderation approach focused on the word "kill" would flag nearly every person playing a game where there are two opposing teams fighting for the same goal. However, ToxMod is able to assess tone, pitch, and context in order to differentiate between "I am going to kill you…before you capture my flag" as compared to "I am going to kill you" in the context of a threat to someone's life.

Data Collection

Data was collected via Modulate's ToxMod, a voice moderation solution which analyzes online speech for emotion, volume, transcribed content and intention, and other related signals in order to identify harmful or maliciously-intended content. ToxMod runs in a variety of games and online social platforms. For this project, three indie studios (each with a few hundred-thousand active monthly users) who enlist ToxMod within their communities consented to the anonymized data from their platforms being used for the purpose of better understanding the landscape of toxic online interactions.

For this analysis, Modulate selected data that was collected within a preset 30 day time period[1]. In total, voice data was collected and analyzed for 11,462 players.

Perceived adult versus perceived underage players

In addition to flagging offensive content, ToxMod can enlist a machine learning model to determine the perceived age of a player. Specifically, it can determine if a player is prepubescent (perceived to be underage) or post-pubescent (perceived to be an adult)[2].

Based on the perceived age confidence of ToxMod, Modulate selected post pubescent players who were active on consenting titles within the target time frame, resulting in a population of 5,731 likely-adult players. Modulate then uniformly randomly sampled prepubescent players to obtain an equally sized population of likely-underage players to compare against. In total, voice over data was collected and analyzed for 11,462 players. Of these, 50% (5,731) were in the perceived underage (non-adult) group, and the other half were in the perceived adult (post-puberty) group.

Voice-chat offenses

The primary function of ToxMod is to identify potentially harmful voice content by examining content, emotional nuance, and other auditory characteristics of the clip and surrounding speech from the speaker and others in the conversation. These signals are fed into ToxMod's machine learning models which ultimately produce a classification of the primary type of harm occurring in the clip. ToxMod can identify a range of harmful behaviors including adult language, sexual vulgarity, violent speech, audio assault (i.e., voice raids), gender/sexual hate speech, racial/cultural hate speech, and "other harmful speech". These categories are outlined in Table 1.

It is important to note that the categories of sexual vulgarity and racial/cultural hate speech have several subcategories within them. Sexual vulgarity is broadly defined as any graphic description of a sexual act or sexualized body part. This includes sexual vocabulary, propositioning, and history survey. Racial/cultural hate speech broadly refers to any actions which show disrespect or malevolence towards another player or demographic group for reasons relating to their cultural heritage or racial/ethnic background. This includes racial hate[3], cultural hate, political hate, and religious hate (see Table 1).

[1] Modulate does not retain data for longer than a 30 day period.

[2] The model is a binary classifier that uses mel-frequency cepstral coefficients as input features. It outputs a score between 0 and 1, where 0 represents the post-pubescent class and 1 represents the prepubescent class. Thus, the model's output represents its confidence that a given clip contains a perceived underage player. Due to the inherent uncertainty of the model on clips where the output is around 0.31-0.69, Modulate internally defines users who are scored below a 0.3 as a perceived adult and those scored above 0.7 as a perceived underage player. Using these thresholds, players were grouped and labeled as perceived adult or perceived underage players. The analysis presented is done based on those labels. It is also worth noting that the model is tuned to minimize any bias in age prediction across gender.

[3] Use of the "n-word", in most groups, is an example of this type of offense. Modulate does make sure to take note of the group that's having the discussion though. For example, in the case of the n word, some black players have reclaimed the slur as a positive part of their culture, and over-aggressively banning any usage of the term actually adversely impacts these already underserved communities and makes them feel less welcome on the platform.

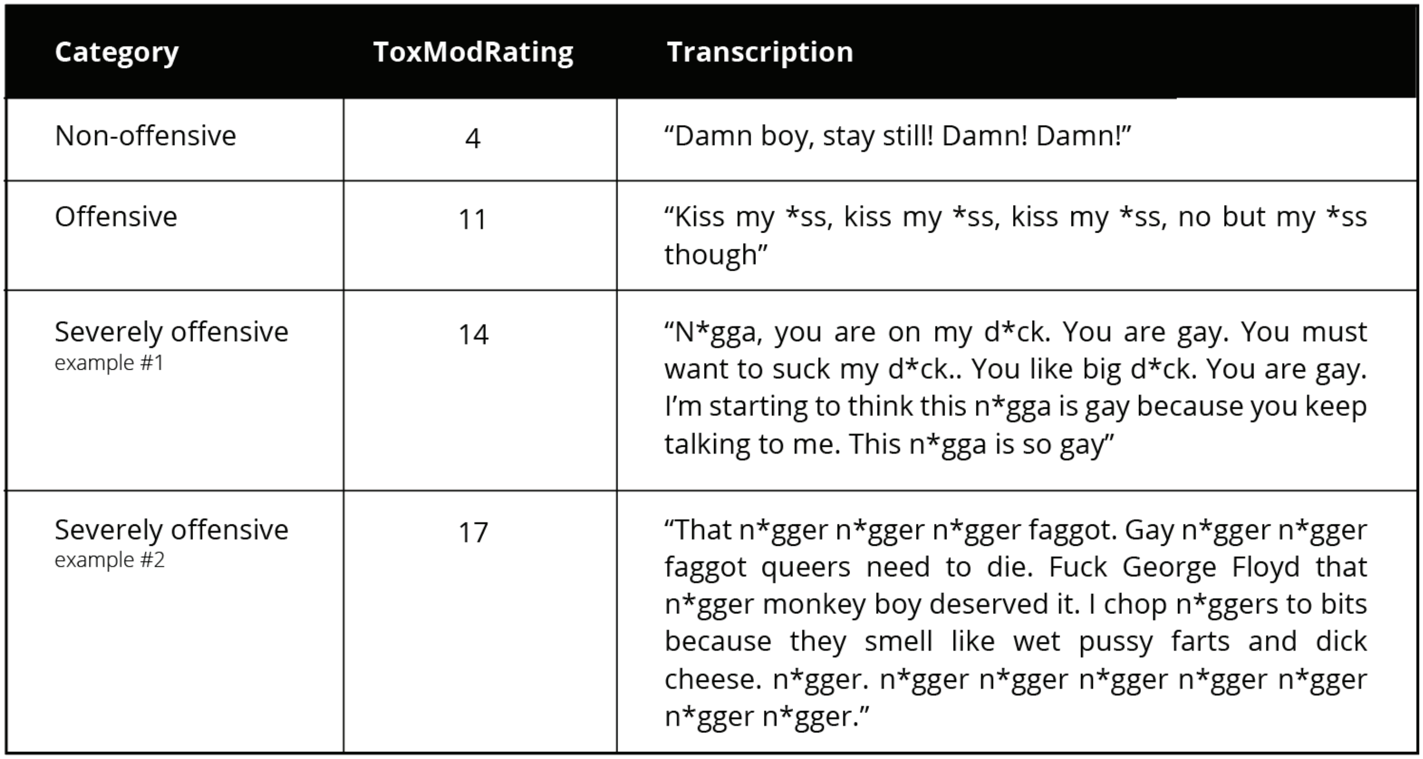
Table 1. Definitions and examples of offensive speech identified and moderated by ToxMod

	Definition	Example
Adult language	Any use of terminology deemed problematic in the presence of underage participants, or otherwise simply deemed universally undesirable within "respectable" company.	F*ck, sh*t, etc
Sexual vulgarity	Any graphic description of a sexual act or sexualized body part	
Sexual vocabulary	Terminology relating to one's biology or sexual acts	Penis, vagina, f*ck*ng, cumming
Propositioning	Requesting, soliciting, or demanding sexual behaviors	"Wanna f*ck?", "Suck my d*ck"
History survey	Asking intrusive or offensive questions about sexual history	"Are you a virgin?"
Violent speech	Speech acts designed to make another player feel physically unsafe.	"Better lock your doors tonight, cause I'm going to be hunting you down."
Audio assault	Production of loud, repetitive, or otherwise intrusive noises that dominate the voice channel and prevent conversation.	Using a soundboard to play a fart sound over and over when someone tries to speak
Gender / sexual hate speech	Any derogatory speech targeting someone's gender identity or sexual orientation.	"girls suck at games", using "queer" or "gay" as an insult
Racial / cultural hate speech	Any actions which show disrespect or malevolence towards another player or demographic group for reasons relating to their cultural heritage or racial/ethnic background	
Racial hate	Use of racial epithet	N-word as an insult
Cultural hate	Insinuations of cultural superiority or inferiority	Speaking in dismissive terms about the "ghetto"
Political hate	Any direct reference to political parties or affiliations as well as making clearly inflammatory remarks	Claims that Trump or Biden supporters deserve to die
Religious hate	Insult others based on their religious affiliations or identity or intentionally inflammatory comments about the existence of religious deities, the observance of religious traditions and rituals, or the avoidance of religious taboos	"F*ck*ng atheists"
None	None is selected if ToxMod did not identify harmful behavior of any kind within the given clip considering the content, context, and emotional nuance of the content.	

Offensive and severely offensive content

In addition to determining the category of offense, ToxMod produces a numerical score between 0 and 20 indicating the severity of the harmful behavior. A score is generated for every voice clip that ToxMod flags as offensive. Any clip which ToxMod scores above a 10 is considered offensive, with clips above a score of 13 being considered severe offensive. It is important to note that offenses that score above 10 may not be ban-worthy by the game's code of conduct but are still considered offensive in the listed primary category of "non-severe offenses." An offense is flagged as a severe offense if ToxMod ascribes it a score of 13 or higher. A severe offense is generally considered a bannable offense from moderators. Transcribed examples from the data for non-offensive, offensive, severely offensive offenses can be found in Table 2.

Table 2. Transcribed audio clips across ToxMod severity categories



Category	ToxModRating	Transcription
Non-offensive	4	"Damn boy, stay still! Damn! Damn!"
Offensive	11	"Kiss my *ss, kiss my *ss, kiss my *ss, no but my *ss though"
Severely offensive example #1	14	"N*gga, you are on my d*ck. You are gay. You must want to suck my d*ck.. You like big d*ck. You are gay. I'm starting to think this n*gga is gay because you keep talking to me. This n*gga is so gay"
Severely offensive example #2	17	"That n*gger n*gger n*gger faggot. Gay n*gger n*gger faggot queers need to die. Fuck George Floyd that n*gger monkey boy deserved it. I chop n*ggers to bits because they smell like wet pussy farts and dick cheese. n*gger. n*gger n*gger n*gger n*gger n*gger n*gger n*gger."

ToxMod's scoring models are calibrated against the decisions of real moderators in order to ensure accuracy. For the categories considered in this analysis, ToxMod's classification has been >99% accurate for severe offenses in all relevant titles for at least two months[4].

[4] Accuracy is determined by customer reviews of ToxMod reports. For the severe offenses flagged by ToxMod, customers have chosen to act on 99% of them.

Results

Overall prevalence and type of offenses

In total, 25,291 offenses were flagged within the data set. 26.43% of players had at least one incidence of offensive language. Of all the players, 21.39% had only non-severe offenses whereas 5.03% of them had logged at least one severe offense. As can be seen in Table 3, racial/cultural hate speech and sexual vulgarity were the most common offenses flagged by ToxMod.

Table 3. Percentage of all offenses by all users by category and severity

	Severe	Non-severe	Total offenses
Racial/cultural hate speech	50.45%	53.47%	53.35%
Sexual vulgarity	42.53%	32.83%	33.21%
Gender/sexual hate speech	6.90%	12.64%	12.42%
Other	0.10%	1.06%	1.02%
Total offenses	985	24,306	25, 291

More than half (53.35%) of all offenses were rated by ToxMod to have the primary category of 'racial cultural hate speech'. Sexual vulgarity was the second most common offense (33.21%), followed by gendered sexual hate speech (12.42%). Sexual vulgarity was the only category to be rated by ToxMod as occurring more often as a severe, than non-severe, offense.

Perceived child versus perceived adult offenses

Perceived adult players were more likely to have at least one offense (36.28%) than perceived underage players (16.58%), including non-severe (28.98 and 13.8%, respectively) and severe (7.29% and 2.77%, respectively) offenses. This can be seen in Table 4.

Table 4. Total percentage of participants with at least one offense as flagged by ToxMod

	Percevied underage	Perceived adult	Total offenses
At least one offense	16.58%	36.28%	26.43%
Only non-severe offenses	13.8%	28.98%	21.39%
At least 1 severe offense	2.77%	7.29%	5.03%

Offenses were also evaluated in regards to their category and perceived age. As seen in Table 5, racial/cultural hate speech and sexual vulgarity were the most common offenses flagged by ToxMod for both perceived adult and underage players.

Table 5. Percentage of all offenses by all users by category and age

Category of offense	Percevied underage	Perceived adult
Racial/cultural hate speech	41.99%	55.63%
Sexual vulgarity	38.17%	32.22%
Gender/sexual hate speech	18.23%	11.25%
Other hate speech	1.61%	0.91%

Recidivism of offenders: Sexual vulgarity and racial/cultural hate speech

Additional analyses were undertaken to assess whether certain offenses might be a so-called 'gateway' behavior to more frequent offenses. That is, additional analyses were undertaken to determine whether players exhibiting specific behaviors would subsequently become more likely to misbehave in a broader range of categories. In order to explore this hypothesis, recidivism rates were examined among severe and non-severe offenders, across age categories, within the two most common offense categories: racial/cultural hate speech and sexual vulgarity.

Racial/Cultural hate speech offenders

Among non-severe offenders, perceived underage players with hate speech offenses were likely to have an average of 2 total offenses (across type) within a 30 day period. Perceived underage players flagged with severe vulgar offenses were found to have an average of 3 - 5 total offenses (across type) within a 30 day period.

Perceived adult players flagged with only non-severe vulgarity offenses were found to have an average of 3 - 6 total offenses (across type) within a 30 day period. Those flagged with severe vulgar offenses were likely to have an average of 6 - 27 offenses (across type) within an average of 30 days.

These outcomes are displayed in Table 6.

Table 6. Average number of offenses for perceived underage and perceived adult players flagged with hate speech offenses over a 30 day measurement period

	Percevied underage	Perceived adult	Average
All non-severe offenders	5.18	10.42	8.93
Vulgarity	1.50	3.38	2.84
Hate speech	2.59	5.73	4.84
Severe offenders	12.81	36.97	30.63
Vulgarity	2.06	6.40	5.26
Hate speech	9.81	27.76	23.05

Vulgarity offenders

Among non-severe offenders, perceived underage players with vulgarity offenses were likely to have an average of 2 - 3 total offenses (across type) within a 30 day period. Among severe offenders, perceived underage players were found to have an average of 3 - 5 total offenses (across type) within a 30 day period.

Perceived adult players flagged with only non-severe vulgarity offenses were found to have an average of 3 - 6 total offenses (across type) within a 30 day period. Those flagged with severe vulgar offenses were likely to have an average of 14 - 26 offenses (across type) within an average of 30 days.

These outcomes are displayed in Table 7.

Table 7. Average number of vulgarity, hate speech, and total offenses for perceived underage and perceived adult players flagged with vulgarity offenses over a 30 day measurement period

	Percevied underage	Perceived adult	Average
All non-severe offenders	5.31	9.00	791
Vulgarity	2.39	3.35	3.06
Hate speech	1.92	4.55	3.78
Severe offenders	10.08	40.39	31.99
Vulgarity	5.70	14.05	11.73
Hate speech	2.90	22.20	16.85

Discussion

Content moderation has been an area of continual learning and growth within the video game industry. However, this has primarily been limited to the examination and moderation of text-based chat. In the current work, Take This partnered with Modulate to generate a better understanding of the verbal communication in games to better understand moderation needs within this space.

Results indicated that there are high levels of voice-chat offenses within gaming spaces. Racial/cultural hate speech constituted more than half of all offenses by all users by category and was more likely to be non-severe than a severe offense. Gender/sexual hate speech was the only other specific category found to be more likely to be non-severe than a severe offense. Together, this suggests that these specific kinds of hateful language have become embedded within the normal speech patterns in this environment and indicate a normalization of this specific kind of speech within gaming cultures. This is in line with previous research that has found evidence for a normalization of extreme language in the text-chat of gaming spaces, specifically in regards to hateful language (Kowert et al., 2022). Notably, perceived adult players were more likely to be flagged for an offense than perceived underage players (36% and 17%, respectively) and twice as likely as perceived underage players to have at least one severe offense.

Sexual vulgarity was the only type of offense that was more likely to be severe than non-severe, speaking to the potential accelerated trajectory of this kind of behavior. Analyses also revealed that those with severe vulgarity offenses were more likely to commit other offenses of any kind, particularly among perceived adult players. Further research is needed to explore if the use of this specific kind of language is a potential gateway or tipping point to more frequent offensive behavior in gaming spaces.

While these findings provide a much needed, foundational step in understanding the landscape of voice chat within gaming spaces, we caution against their broad interpretation. It is possible that any of these findings discussed within this paper could be, at least partially, shaped by ToxMod's design. For example, it is possible that ToxMod rates sexual vulgarity offenses as more severe, as compared to other types of offenses, whereas traditional human moderators may not. While there is no reason to believe this is the case, it is not possible to rule this out due to the subjective nature of determining the level of offensiveness of explicit language.

Conclusion

Taken together, this work supports the notion that there is a high prevalence of severe language and hateful sentiments within the social environment of digital gaming spaces. Knowing that hate speech is a prevalent occurrence within voice chat in addition to text-based chat in gaming spaces highlights the urgent need for an expansion of moderation efforts within this space. When a culture - any culture - is filled with hateful sentiments, individuals can become desensitized to hate speech and, over time, it can foster polarization between communities and increase biases among community members. When hate is allowed to spread without consequence, it normalizes hate in all spaces. This so-called "normalization of hate" within gaming spaces is a problematic and dangerous element for game makers and players alike.

The results of this research also point to an important aspect of what is often referred to in shorthand as "toxic gamer culture". Specifically, it reveals that the social environment of digital gaming spaces is characterized by hateful language, both racial/cultural and gender/sexual, and that these offenses constitute the most frequent incidences within these spaces. Understanding that "toxicity" within gaming spaces does not only refer to the prevalence of trash talking or griefing, but rather the pervasiveness of hate speech is an important distinction when discussing moderation strategies. More effective moderation, particularly within voice chat, needs be a top priority among game makers and the prevalence of hateful language should alarm educators, law enforcement, anti-terrorism experts, and others working to reduce hate and violence in both online and offline environments.

References

Anti-Defamation League. (2022). Hate is No Game: Harassment and Positive Social Experiences in Online Games 2021.

Consalvo, M. (2012). Confronting toxic gamer culture: A challenge for feminist game studies scholars.

Kowert, R., Botelho, A., & Newhouse, A. (2022). Breaking the Building Blocks of Hate: A case study of Minecraft servers. A report from the Anti-Defamation League. (ADL) center of Technology and Society.

Kowert, R., & Cook, C. (2022). The toxicity of our (virtual) cities: Prevalence of dark participation in games and perceived effectiveness of reporting tools. Proceedings of the 55th Hawaii International Conference on System Sciences.

Paul, C. A. (2018). The toxic meritocracy of video games: Why gaming culture is the worst. U of Minnesota Press.

Ross, A. (2021, July 20). GDC 2021: Research on online video game voice chat has come a long way. Massively Overpowered. Retrieved from https://massivelyop.com/2021/07/20/gdc-2021-research-on-online-video-game-voice-chat-has-come-a-long-way/

Subspace. (2021). Why in-game voice chat became essential to multiplayer gaming and how it impacts the metaverse. Subspace. Retrieved from https://subspace.com/resources/in-game-voice-chat-and-the-metaverse

Tencent Cloud. (2022, Jan 7). The power of in-game voice chat. Hackernoon. Retrieved from https://hackernoon.com/why-every-multiplayer-game-needs-in-game-voice-chat

Williams, S. (2015, August 17). VoIP and gaming go hand in hand. IT Brief. Retrieved from https://itbrief.co.nz/story/voip-and-gaming-go-hand-hand