A Framework for Analysis of Speech and Chat Content in YouTube and Twitch Streams

Steven Coats
English, Faculty of Humanities, University of Oulu, Finland
E-mail: steven.coats@oulu.fi

Abstract

Online streaming platforms have become important sites of interaction and communication, but relatively little research into streaming platforms has considered the combined discourse of speech transcripts and live chat streams. In this paper we describe a pipeline approach that can integrate speech transcripts with live chat content in order to create structured documents from streams recorded on the platforms YouTube and Twitch. Built on common streaming protocols and the open-source Python library yt-dlp, the notebook comprises modular script components for data download and organization of transcripts and live chat and can additionally retrieve audio, video, and other streamed content. Additional pipeline modules can be used for automatic speech-to-text transcription of the video stream and incorporation of models for specific analytical tasks such as automatic video classification, gesture identification, or facial recognition. The paper demonstrates use of the notebook to output a time-stamped, structured combined speech/chat html file and proposes two possible analyses: consideration of chat density, and zero-shot classification of video content.

Keywords: streaming, YouTube, Twitch, multimedia, multimodal analysis, speech-to-text

1. Introduction

The increase in popularity of online streaming in the past 15 years has given rise to new, complex computer-mediated communication (CMC) environments which combine video, audio, written text, and graphical images, among other elements (Sjöblom et al. 2019). With the standardization of technical protocols for streaming and increased access to bandwidth, memory, and storage since the late 2000s, live streaming (and sharing of recordings of live streams) has become a common CMC modality on a variety of platforms which may be specialized for stream content types such as gaming and esports, talk and discussion, or other content. The most widely used streaming platforms are YouTube, which hosts a variety of streaming content, and Twitch, whose focus is primarily on gaming and esports. From the perspective of corpus-based studies of language and interaction, streams can be environments which embody multimedia and multimodal communication at multiple levels: the speech and visual content (e.g. facial expressions or gestures) of the streamer on camera, as well as, potentially, those of other persons physically or present in the same environment; the text, speech and visual content of other streams captured in the video output (in the case of Twitch, this often includes a screen showing gameplay); the text and graphical image content of the accompanying live chat, which can have hundreds or thousands of participants; and the text and graphical content of system messages such as donations or tips to the streamer, among others.

Although streaming has become a popular (and economically important) form of CMC, and a substantial research literature, particularly in computer science, has considered aspects of streaming, relatively few studies have been based on multimodal corpora which record the multiple communicative levels present in streams. High-speed chats in live streams have attracted research attention, but few studies have compared the content of live chat with the speech content of the streamer as represented in automatic speech recognition (ASR) transcripts.[1] Likewise, corpus-based comparisons of chat, transcript content, and visual or auditory content remain few.

While the modalities of this kind of interactive environment have been described in the context of CMC research, many studies have focused on disentangling the potentially complex configuration of interlocutor dynamics from a theoretical perspective, for example by describing the basic functionality of massive anonymous chat environments or analyzing aspects of online game streaming from ethnological and sociocultural perspectives. Empirical, corpus-based studies which compare the speech of the video stream, the text content of the chat, the graphicons used by participants, the automated system messages, and the content of the video stream, *inter alia*, have been relatively few, particularly from a corpus-linguistic framework. In part, this is due to the complex nature of the underlying multimodal data, which comprises a variety of video, audio, text content in different formats, which can be difficult to work with.

In this study we provide a preliminary script pipeline for capturing and combining recorded stream audio transcripts with live chat content in a timestamped tabular format that can serve as the starting point for corpus-based analysis.[2] The framework, in a notebook environment accessible via Google's Colab cloud computing environment, can also be used to collect video content. Further developments of the framework will incorporate models for various types of audiovisual analysis.

In the next section, some previous research on live streaming is presented. Section 3 describes the main elements of the pipeline and shows an excerpt from a combined speech transcript-live chat output file. In Section 4, use cases are noted: a consideration of chat density, and the potential for automated video content classification analysis. The study concludes with a brief outlook for future developments with the pipeline.

[1] See Coats (2024) for a comparison of ASR transcript content with video comments.

[2] https://t.ly/le6_e

2. Previous research

Live streams can have hundreds or thousands of participants; chat windows in live streams can therefore be fast paced and often lack interactional coherence. Herring (1999) noted that elements such as simultaneous feedback and turn adjacency can be lacking in chat environments with a large number of participants, which, however may offer possibilities for heightened interactivity and language play.

Hamilton et al. (2014) undertook an ethnographic study of Twitch gaming communities, proposing that community identity can coalesce around shared experiences in gaming streams, including in live chats. They proposed, however, that participants in massive streams with more than 1,000 viewers are focused mostly on the activities of the streamer, rather than on community interaction, which at this scale is subject to "breakdowns" due to its relative incoherence.

Ford et al. (2017) compared Twitch live chat samples from massive chats with 10,000 or more participants to samples from smaller chats, with 2,000 or fewer viewers. They found that larger chats tend to have shorter messages and more repeated content, often in the form of emotes (i.e. customized graphicon images rendered inline with chat text). Despite its seeming incoherence, "crowdspeak" can serve to consolidate in-group collective identities, for example through use of emotes or lexical items specific to a community. Harpstead et al. (2019) surveyed published research into online game streaming. They found that although many studies have been published, several desiderate remain, including "investigations that make use of broader interaction data" (p. 116).

Corpus-based studies of the language of game streaming platforms have been relatively few. Olejniczak (2015) conducted a study on a 17,500-word corpus of chat content manually collected from Twitch, finding that streams with larger numbers of participants typically have shorter chat messages. Kim et al. (2022) compiled a corpus of 15m words from Twitch stream chats to examine emotes, finding that "toxic emotes" which are used to express negative or derogatory content can be challenging to detect automatically.[3] Emotes can be used, for example, to bypass word filters or stoke racial resentments.

Recktenwald (2017) proposed a columnar transcription scheme for the analysis of the multiple communicative configurations possible in a gaming live stream: one column records timestamps, a second the speech of the streamer, the third column describes events within the game, and the fourth records chat comments. While Recktenwald used the scheme for manual transcription of speech content, this basic layout for a combined transcript is exemplified by the files generated by the automated method of the pipeline introduced in this study.

Streaming platforms have also become important economically. Streams, and recordings thereof, can be monetized by the platforms on which they are hosted. Many streamers accept donations or tips in a stream, link to paid services or e-commerce sites, or offer other kinds of paid content (Zhou et al. 2019). Johnson and Woodcock (2019) considered the economic and sociocultural implications of livestreaming of games, especially for the gaming industry itself. Yu et al. (2018) analyzed the relationship between in-stream engagement and viewer donations or gift-giving. There are few corpus-based studies of transactional events within CMC contexts, however.

In general, while a great many studies have considered aspects of streaming, the majority focus on technical considerations or larger sociocultural issues. Corpus-based studies of speech, discourse, and interaction in streams, particularly in the sense of multimodal activity, remain relatively few, in part due to the challenges inherent to wrangling the data into formats amenable to corpus analysis. Several tools exist for harvesting video and chat data from Twitch and YouTube.[4] These libraries can retrieve the JSON file of a recorded stream's live chat and render the data in various formats, but for the most part do not retrieve speech transcripts from those streams. In the following section, we describe a notebook-based pipeline for collection of chat, speech transcripts, and other data from streams.

3. Data collection pipeline

YouTube and Twitch steams are not equivalent in terms of the affordances available to streamers and viewers or the data that can be retrieved by researchers. Nevertheless, the basic structure is similar for the two platforms, and the use of common technical protocols means that the main content (video, audio, live chat) can be retrieved using functions from open-source libraries. The pipeline in this study uses yt-dlp, a Python library for the retrieval of streamed content. The steps are depicted in Figure 1.

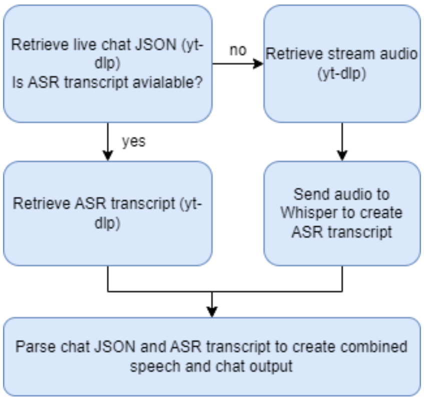

Retrieve live chat JSON (yt-dlp)
Is ASR transcript avialable?
no
Retrieve stream audio (yt-dlp)
yes
Retrieve ASR transcript (yt-dlp)
Send audio to Whisper to create ASR transcript
Parse chat JSON and ASR transcript to create combined speech and chat output

Figure 1: Flowchart of pipeline

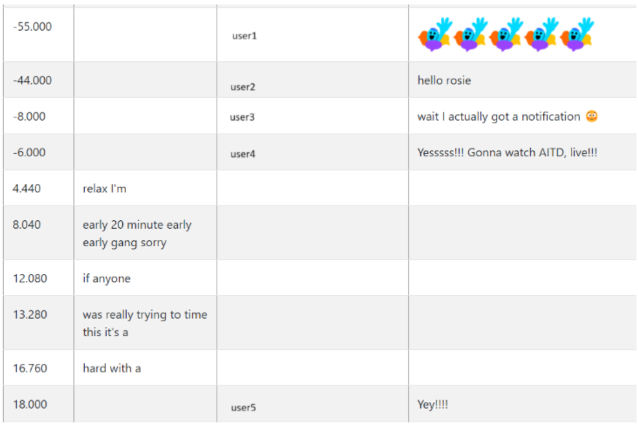
Yt-dlp is used to retrieve the content of the live chat stream as a JSON file.[5] For streams for which an ASR transcript

[3] The authors provide examples in which an emote depicting a bucket of fried chicken, the "KFC emote", is used to make derogatory reference to racial/ethnic identity.

[4] E.g. https://github.com/lay295/TwitchDownloader, https://github.com/xenova/chat-downloader

[5] As of April 2024, yt-dlp can retrieve YouTube live chat

is available (usually most recorded videos on YouTube), the pipeline retrieves the transcript in the default VTT file format. For recorded streams for which no ASR transcript is available, the pipeline records the audio of the stream and feeds it to Whisper (Radford et al. 2022) for generation of an ASR transcript. The live chat JSON and the ASR transcript files are then parsed and the speech and chat elements rendered in a data frame in the correct order. The output is saved as an HTML file. Figure 2 depicts an excerpt from an output file, a recorded stream of the popular YouTube personality PewDiePie. The first column shows the timestamp for the utterance or chat contribution. Chatrooms can be opened prior to the start of the stream, which is why some chat entries have negative timestamps. The second column shows the transcript of the speech in the video – in this case, the speech of PewDiePie.[6] The third column shows usernames (anonymized in this screenshot), and the fourth column the chat message. The pipeline renders standard emoji (in the third row) as well as custom, non-Unicode emoji which are used in a particular channel (the emoji in the first row).



-55.000		user1	
-44.000		user2	hello rosie
-8.000		user3	wait I actually got a notification 😳
-6.000		user4	Yesssss!!! Gonna watch AITD, live!!!
4.440	relax I'm		
8.040	early 20 minute early early gang sorry		
12.080	if anyone		
13.280	was really trying to time this it's a		
16.760	hard with a		
18.000		user5	Yey!!!!

Figure 2: Excerpt from output file

4. Use cases

An aligned transcript containing speech and chat messages can be used as the basis for investigations of the properties of multimodal CMC, including analyses of grammatical phenomena, lexis and discourse, and emoji and emotes. In addition, information from the aligned transcript can be utilized in the context of analytical steps undertaken on the underlying video data.

4.1 Chat density

One possible approach is to analyze chat density in streams and correlate stream content and/or speech messages with periods of high or low chat density. Figure 3, again from the PewDiePie stream cUUuRK3Rm4k, shows density of streamer speech and live chat messages. In this figure, the blue line represents the density of chat messages per minute, and the orange line the number of utterances per minute by the streamer. As can be seen, there is little communication in the chat stream prior to the streamer starting the stream (at 0 minutes). Thereafter, densities remain fairly constant: speech at approximately 20-25 utterances per minute, and chat between 25 and 50 messages per minute. The exception is from minute 105 until the end of the stream. Here, examining the aligned transcript provides clues: The large increase in number of chat messages is prompted by the streamer interacting directly with his audience by saying "I totally forgot there were so many people watching, hey it's good to see you guys", prompting a large number of messages and responses in the chat; this is followed by several more comments by the streamer addressing the audience directly.

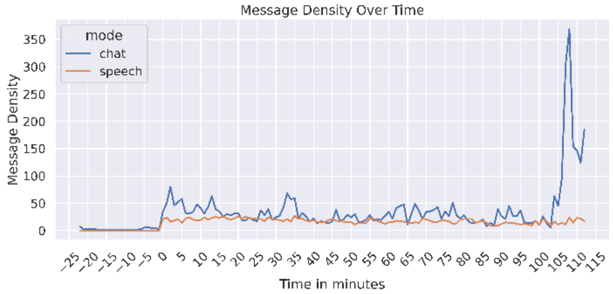


Figure 3: Chat and speech density for stream cUUuRK3Rm4k

4.2. Automated analysis of video streams

The notebook format of the pipeline may be suitable for combining text-based analysis of speech and chat messages with automated methods for the analysis of multimodal visual-textual content, for example by importing models from Huggingface. Stream segments with high levels of chat activity can be analyzed with models for zero-shot classification of visual content (Ni et al. 2022). Using individual speech turns and/or chat messages as the input texts in a classification task may provide insight into the dynamics of the underlying multimodal communication by clarifying who is commenting on the video and who is commenting on (for example) other chat or speech content. The incorporation of automated video analysis components into the Colab environment for the framework is planned.

5. Summary and outlook

CMC in the past 30 years has undergone a shift from primarily text-based modalities such as message boards or chatrooms towards multimedia environments in which video, audio, text, and images are all shared in real-time in streams. These complex environments require new methods for the organization and curation of data in corpus formats that are suitable for a variety of analytical approaches. The stream pipeline described in this paper presents a notebook environment which retrieves streamed data and combines speech transcripts with live chat messages, allowing the analysis of discourse and graphical content which prompt high rates of chat density. In addition,

streams by default, but to retrieve live chat streams from Twitch video on demand, a patch must be installed (https://github.com/yt-dlp/yt-dlp/pull/1551).

[6] https://www.youtube.com/watch?v=cUUuRK3Rm4k

the environment can integrate tools for ASR and video analysis which allow multimodal comparisons to be undertaken.

Future developments with the pipeline will have three main focuses: First, to consider streams not only from YouTube and Twitch, but from other platforms, covering different kinds of content. Second, to incorporate versatile video analysis models which can provide automatic descriptions of video content such as in-game developments. Third, to incorporate video recognition models that can account for facial expressions and gestures (Parian-Scherb et al. 2022). Models for automatic video content recognition can then be used to generate outputs which can be analyzed in the context of a stream's speech, chat, or system message content.

Advances in AI models for the analysis of video data are ongoing, and in the immediate future, automated annotation of streaming data will undoubtedly be feasible. A notebook-based data collection and analytical environment such as the one presented in this study, which allows time-stamped transcripts of speech and chat content to be combined, will provide a foundation for further developments in the corpus-based analysis and understanding of multimodal online interaction.

6. References

Coats, S. (2024). Commenting on local politics: An analysis of YouTube video comments for local government videos. *Research in Corpus Linguistics*.

Danesi, M. (2017). *The semiotics of emoji: The rise of visual language in the age of the internet*. Bloomsbury.

Ford, C., Gardner, D., Horgan, L. E., Liu, C., Tsaasan, A. M., Nardi, B., & Rickman, J. (2017). Chat speed OP PogChamp: Practices of coherence in massive Twitch chat. In *CHI EA '17: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems May 2017* (pp. 858–871). https://doi.org/10.1145/3027063.3052765

Hamilton, W. A., Garretson, O., & Kerne, A. (2014). Streaming on Twitch: fostering participatory communities of play within live mixed media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1315–1324).

Harpstead, E., Rios, J. S., Seering, J., & Hammer, J. (2019). Toward a Twitch research toolkit: A systematic review of approaches to research on game Streaming. In *Proceedings of the Annual Symposium on Computer-human Interaction in Play* (pp. 111–119).

Herring, S. (1999). Interactional coherence in CMC. *Journal of Computer-Mediated-Communication*, 4(4). https://doi.org/10.1111/j.1083-6101.1999.tb00106.x

Johnson, M. R., & Woodcock, J. (2019). The impacts of live streaming and Twitch.tv on the video game industry. *Media, Culture & Society*, 41(5), 670–688. https://doi.org/10.1177/0163443718818363

Kim, J., Wohn, D. Y., & Cha, M. (2022). Understanding and identifying the use of emotes in toxic chat on Twitch. *Online Social Networks and Media*, 27. https://doi.org/10.1016/j.osnem.2021.100180

Konrad, A., Herring, S. C., & Choi, D. (2020). Sticker and emoji use in Facebook Messenger: Implications for graphicon change. *Journal of Computer-Mediated Communication*, 25(3), 217–235. https://doi.org/10.1093/jcmc/zmaa003

Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., & Ling, H. (2022). Expanding language-image pretrained models for general video recognition. *arXiv*, cs.CV, 2208.02816. https://doi.org/10.48550/arXiv.2208.02816

Olejniczak, J. (2015). A linguistic study of language variety used on twitch.tv: Descriptive and corpus-based approaches. In: *Proceedings of RCIC'15: Redefining Community in Intercultural Context, Brasov, 21–23 May 2015* (pp. 329–334).

Parian-Scherb, M., Uhrig, P., Rossetto, L., Dupont, S., & Schuldt, H. (2023). Gesture retrieval and its application to the study of multimodal communication. *International Journal on Digital Libraries*. https://doi.org/10.1007/s00799-023-00367-0

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv*:2212.04356 [eess.AS]. https://doi.org/10.48550/arXiv.2212.04356

Recktenwald, D. (2017). Toward a transcription and analysis of live streaming on Twitch. *Journal of Pragmatics*, 115, 68–81.

Riddick, S., & Shivener, R. (2022). Affective spamming on Twitch: Rhetorics of an emote-only audience in a presidential inauguration livestream. *Computers and Composition*, 64, 102711.

Siever, C. M. (2019). 'Iconographetic Communication' in digital media: Emoji in WhatsApp, Twitter, Instagram, Facebook—From a linguistic perspective. In E. Giannoulis & L. R. A. Wilde, (Eds.), *Emoticons, kaomoji, and emoji: The transformation of communication in the digital age* (pp. 127–147). Routledge. https://doi.org/10.4324/9780429491757

Sjöblom, M., Törhönen, M., Hamari, J., & Macey, J. (2019). The ingredients of Twitch streaming: Affordances of game streams. *Computers in Human Behavior*, 92, 20–28.

Spina, S. (2019). Role of emoticons as structural markers in Twitter interactions. *Discourse Processes*, 56(4), 345–362. https://doi.org/10.1080/0163853X.2018.1510654

Yu, E., Jung, C., Kim, H., & Jung, J. (2018). Impact of viewer engagement on gift-giving in live video streaming. *Telematics and Informatics*, 35(5), 1450–1460.

Zhou, J., Zhou, J., Ding, Y., & Wang, H. (2019). The magic of danmaku: A social interaction perspective of gift sending on live streaming platforms. *Electronic Commerce Research and Applications*, 34, 100815.