AI SPEAK internet video database – VideoBase 1.0 – Collection and characteristics

1. Description

In order to facilitate different research activities that are part of **AI SPEAK project**, and in particular experimental studies aimed at the tasks of lip reading and speech resynthesis from video for Serbian language, **VideoBase 1.0 internet video database** was created, **Table 1**.

VideoBase 1.0 is envisioned as the large scale, structured database of internet videos in Serbian language that have adequate characteristics for subsequent extraction and processing of individual speakers appearing in the scene. It was designed to cover wide range of person appearances and speaker characteristics and provide high quality video and audio streams that are suitable for further analysis and processing. The size of database can be considered as relatively large, in terms of both number of unique audio/video recordings, as well as the number of unique speakers and recording environments.

Due to requirements to have high quality content, all the recordings are limited to studio production and TV broadcast formats. However, such deliberate choice of video sources does not affect the content diversity and its potential applications in the real world settings of uncontrolled recording environments.

Since collected internet videos are protected by copyright, collected video database is mainly intended to be used as internal project resource, available for extensive analyses and testing of developed algorithms and procedures on collected audio-video streams. However, the database also includes additionally generated meta-data about person appearances in the video, which can be made public in the form of anonymized log-files, depending on application. These additional data can be produced by the video analysis tool VideoFace2.0 specifically developed by the project team for the purpose of VideoBase 1.0 downstream applications.

Table 1 – Summary of created AI SPEAK internet video database.

VideoBase 1.0	Characteristics
No. of unique videos	**2400**
Total video duration	**~1363 hours**
Storage size (core video database)	**~850 GB**
Videos duration	~30 minutes
Video streams	1920x1080 @ 25 fps, high bit rate
Audio streams	~100 kbps ABR, compressed
Video container formats	.mp4; .mkv
No. of videos per content provider	600
No. unique speakers	> 100
Audiovisual corpora (talking face videos)	~ 100 hours

2. Characteristics

VideoBase 1.0 core database consists of **2400 unique internet videos with more than 1363 hours of audio-video material.**

All videos are recorded with **high spatial and temporal resolutions of 1920x1080 pixels, and 25 frames per second (fps)**, respectively. Similarly, collected **high fidelity audio recordings** are made with professional production equipment under studio and outdoor recording conditions, and originally encoded for internet audio streaming **with at least 100 kbps or higher data rates**. Audio and video streams are stored as individual containerized video files, in corresponding **audio-video container formats** (e.g. .mp4; .mkv), depending on the specific requirements of the **encoding type** and format characteristics in each case. Such data files are named according to the following naming convention:

```
title␣[duration in mm-ss]␣[identifier]␣[upload date in yyyy-mm-
   dd]␣[video_encoder_ID + audio_encoder_ID].container_format
```

which provides unique identification, recording duration, audio and video encoding type, original title and video publication date.

Duration of videos in VideoBase 1.0 collection mostly **varies between 15 and 30 minutes**, and corresponds to usual TV formats of news broadcasts, reportage interviews in outdoor and indoor spaces, and studio interviews with several guests (discussion talk shows).

Besides **core video database** described **in Table 1**, AI SPEAK internet video database also includes audio-video materials produced by subsequent processing of collected internet video files in VideoBase 1.0. As already mentioned, of particular interest for development of AI models for lip reading and speech resynthesis from video are recordings of speaker faces and mouth regions. Such **structured audiovisual corpora** can be extracted from the original video files by the means of internally developed **VideoFace2.0** video analysis tool and corresponding graphical user interface, which are described in Section 4 and reference [1].

It means that for each original input video file from VideoBase 1.0 collection, additional **meta-data about persons** that are appearing in the video can be produced based on their **face identities**. Such meta data, in the form of corresponding **anonymized log-files**, can be **stored alongside original video files** in the database **for further post-processing**, i.e. downstream tasks. By subsequent loading of original video and its log-file pair, produced meta-data allow for creating different newly generated output videos, e.g. video frames containing only the selected speaker, or the video containing only the person's face or mouth region, including the accompanying audio sequences from the original input video.

More exactly, described post-processing of anonymized log-files associated with video entries in VideoBase 1.0 core database assumes controlled extraction and concatenation of all specific parts of the original input video that correspond to the: 1) selected speaker (unique face identity in the log-file), and 2) defined spatial region of interest (face or mouth region that can dynamically change in the original unconstrained video input). Such video outputs or video stories are sometimes also called **talking face videos**, or **mouth region videos**, and in both cases require that the produced output contains only the selected single speaker with face partially or fully oriented

towards the camera. Thus, developed video analysis tool assigns anonymized person identities based only on the analysis of their face images.

Extraction of 5 minutes of talking face videos from 300 original input video files allows for 25 hours of high quality audio-visual corpora. Since created video database has **600 videos per each of 4 selected content providers** (video productions), it means that the created video collection provides **100-200 hours of high quality audiovisual training data** for planned research tasks.

VideoBase 1.0 collection items are grouped by source (content provider), and include **two types of production formats**. The first type are **regular news broadcasts,** produced by 2 Serbian TV stations with significant market share, while the remaining two sources correspond to the **discussion talk shows** in TV format with several guest speakers. In both cases, some of the journalists and speakers are regularly appearing in all video files from the same source (content provider), which allows for speaker dependent corpora development. At the same time there is also a large variety of other native speakers in each of the videos. In particular, recordings also include short reportages and outdoor recordings in forms of interviews, which are particularly suitable for providing talking face videos from uncontrolled environments.

3. Collection process (VideoCapture)

VideoBase 1.0 creation was supported by software development environment **VideoCapture** that mainly consists of the libraries and tools supporting controlled reproduction and saving of video and audio streams from internet streaming platforms. For more details about used libraries and software frameworks please consult the corresponding configuration files in Appendix A.

In particular, all **2400 high quality internet videos** were recorded by programming scripts that ensure desired level of quality and storage efficiency, **without re-encoding of the originally uploaded content**.

The list of selected 2400 videos with unique identifiers, originally assigned by the internet streaming service, is provided in **Appendix A** in the file “VideoData_list_IDs.txt”.

Videos in the list can be partitioned into four groups corresponding to **4 different content providers** (video productions), with **600 videos from each source**, Figure 1. Duration times of video files, depending on source provider, are summarized in **Table 2**.

Table 2 – Duration time statistics for collected video files depending on content provider.

Video content provider	Video duration [mm:ss]				Total time [hh:mm:ss]
	Minimum	Maximum	Mean	Median	
Dnevnik_RTV	1:37	44:31	25:33	25:40	255:31:54
Dnevnik_N1	25:22	168:53	37:12	34:54	369:01:04
Dobro_Jutro_TANJUG	2:47	86:50	35:26	36:07	354:28:25
Uranak_K1	4:34	81:29	38:26	39:14	384:22:46

In total, **VideoBase 1.0 collection contains 1363 hours of video material** with mean and **median video duration of** 34 and **32 minutes**, respectively. It requires around **850 GB in storage space**.

Besides internet urls and video identifiers, accompanying files listed in **Table A1** of **Appendix A** also contain detailed descriptions of video files (original title, source, and duration), including information about usually available encoding types or formats of collected video and audio streams (descriptions regarding representative examples of encoding types for each content provider are in the corresponding files with prefix "video_enc_" in Appendix A).

Special care and experimentation was made in order to **select optimal encoding types** among the ones available on the streaming platform. An overview of available encoding formats in the case of one randomly chosen sample video from the collection is shown in **Table B1 of Appendix B**.

According to this table, as the preferred encoder for video stream was chosen encoder with **ID 399** ("av01.0.08M.08"), which corresponds to **AOMedia Video 1 (AV1) encoding** standard, with spatial resolution of **1920x1080 @ 25fps**. In this particular case, streaming protocol is DASH (Dynamic Adaptive Streaming over HTTP) with video segments encoded in .mp4 container format. Similarly, as the preferred audio stream encoder was chosen the **MPEG-4 AAC encoder** with Low Complexity profile ("mp4a.40.2"), shown under **ID140** in **Table B1**.

We note that the particular choice of the described audio encoder was driven by the need that audio and video streams are jointly saved in **single** .mp4 **container file format**, which could not be accomplished in the case that some alternative audio encoder like OPUS (e.g. **ID 251**) was used (e.g. in such case compatible container file format would need to be .webm instead of .mp4). However, in the case of both ID 399 and ID 140 the corresponding encoder settings were always chosen to be in the high bit rate range.

When the preferred **ID 399** (**AV1**) **video encoder** was not available, video encoders illustrated in Table B1 were chosen according to the following **predefined ID preference list**: **614** (**VP9**, advanced), **248** (**VP9**, baseline), **137** (**H.264**). Similarly, for the alternative audio codecs the ones with the .m4a container file format were the preferred choice.

(a) (b)

(c) (d)

Figure 1 – Illustration of four selected content providers (video sources) from Table 2: "Dnevnik_RTV" (a), "Dnevnik_N1" (b), "Dobro_Jutro_TANJUG" (c), "Uranak_K1" (d).

4. Developed video analysis tool (VideoFace2.0)

AI SPEAK internet video database is primarily designed to be utilized as a source of audiovisual corpora for training machine learning models. In that sense an accompanying part of the VideoBase 1.0 database are additional research tools that can be used for processing of collected video files and **extraction of specific training data**.

In particular, **VideoFace2.0** is the name of developed video analysis tool for spatial and temporal localization of each unique face in the input video, i.e. face re-identification (ReID) task, which also allows for face cataloging, characterization and creation of structured video outputs for later downstream tasks. Developed near real-time solution, Figure 2, is primarily designed to be utilized in application scenarios involving TV production, media analysis, and as an efficient tool for creating video datasets of talking faces in challenging vision tasks such as lip reading and speech resynthesis from video. Specific elements of designed system and proposed face re-identification procedure are described in **AI SPEAK publication**, **reference [1]**, while the video demonstrations of **generated video stories**, i.e. talking face videos of selected speakers extracted from original input video files, are provided on the corresponding demo page in **reference [2]**.

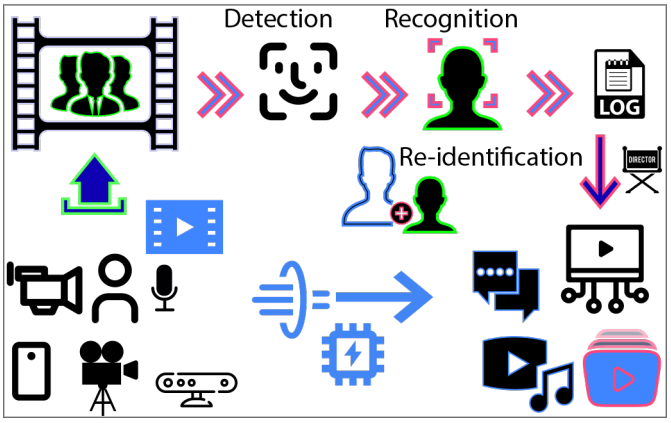


Figure 2 – Processing workflow of developed video analysis tool.

According to **Figure 2**, input video from VideoBase 1.0 is analyzed in search for unique face identities in the foreground (face detection and re-identification tasks under an open set setting), based on which the corresponding **catalogue or log-file of face appearances is made**.

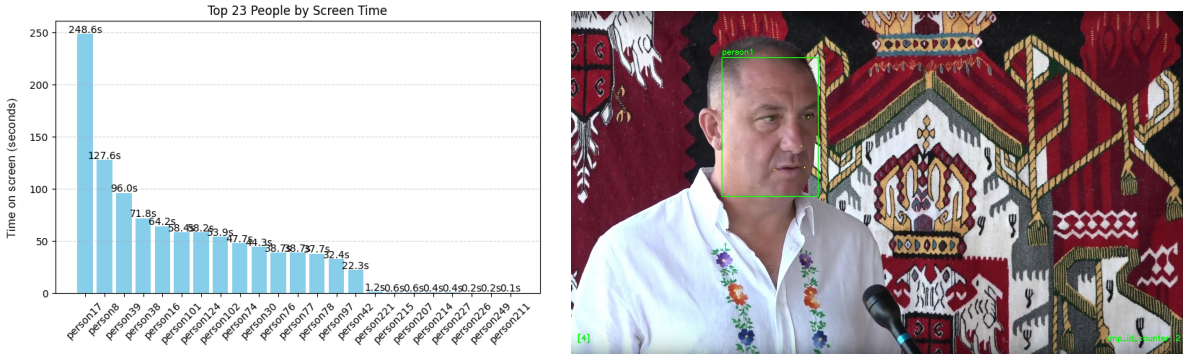
Based on face identity, a unique numeric identifier (**person id**) is assigned to each recognized person in the input video. This allows for **later post-processing** where selected operations are performed only on the input video regions **corresponding to the selected person**.

These can include different statistics, like the on-screen time computation, Figure 3(a), or video compilation (video story) containing only the input video frames with the selected person, Figure 3(b). An example of **generated talking face video**, or video containing only the person's mouth region, are illustrated in **Figure 4**. Specific face regions are extracted based on identified face landmark points, shown over the face image on the left hand side of Figure 4.

Note that the number of unique identifiers assigned to single person in some cases can be more than one, due to system's failure to correctly assign the same identity to different face appearances of the same person (i.e. multiple person ids associated with the same person).

However, such situation, illustrated in **Figure 5**, can be avoided by subsequent selection of desired persons (face based identities) through **designed graphical user interface**, **Figure 6**.

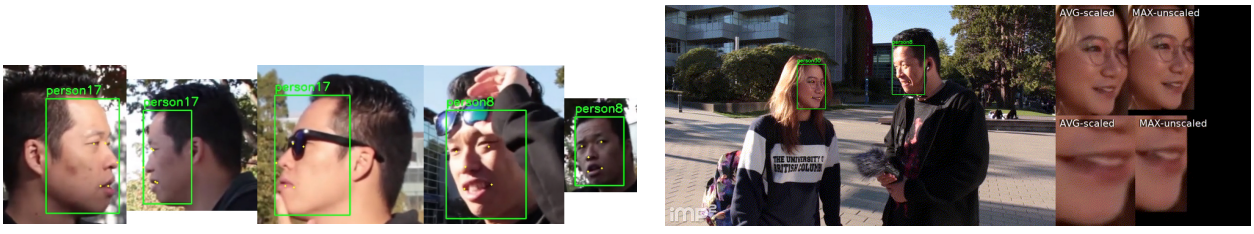
Different options for initial processing of the input video (meta-data, i.e. log-file generation) are shown in **Figure 7**.



(a) (b)

Figure 3 – Post processing of the input video based on generated log-file: on-screen time of persons recognized based on their face identities (a), generated output video with frames containing "person 1" (b).

Figure 4 – Face and mouth region videos (on the right) for selected person on the left. Note that talking face, mouth region videos can be produced with or w/o spatial scaling to average region width or height.



(a) (b)

Figure 5 – Multiple ids associated with the same person due to face appearance variations (a), generated output videos for "Person 30" (b).

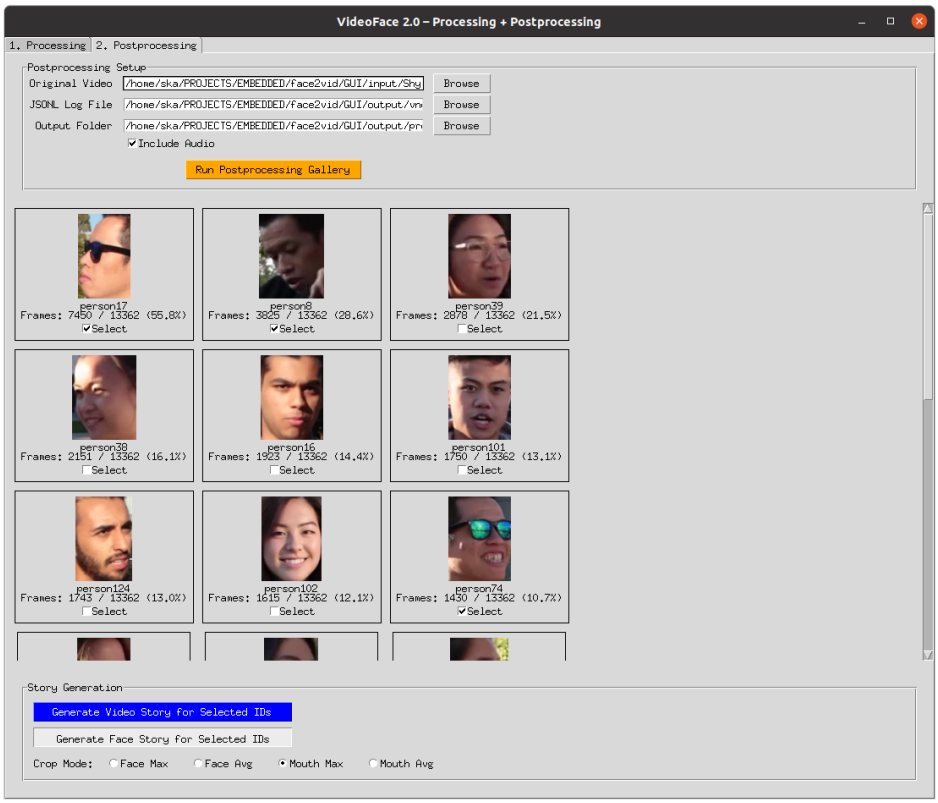


Figure 6 – Graphical user interface for post-processing of input video based on generated log-file.

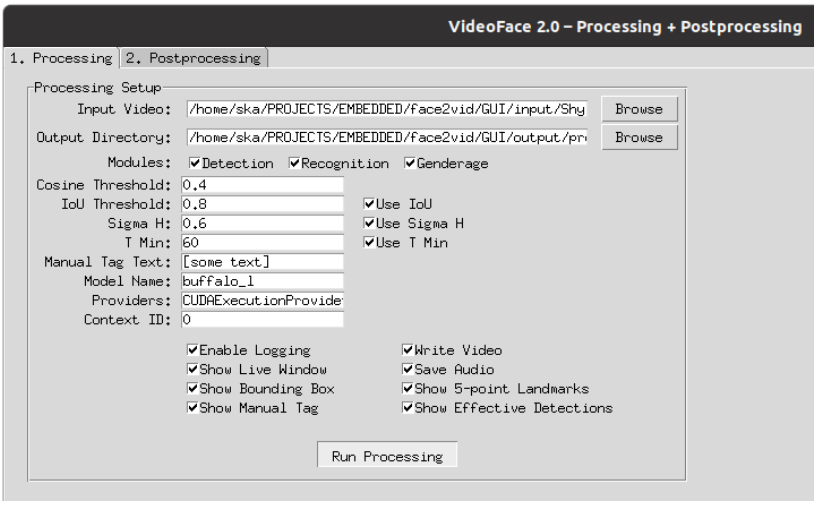


Figure 7 – Initial processing of the input video, options and settings for log-file generation.

References

[1] Branko Brkljač, Vladimir Kalušev, Branislav Popović, Milan Sečujski (**2025**). "Transforming faces into video stories—VideoFace2.0", *14th Mediterranean Conference on Embedded Computing (MECO)*, IEEE, Budva, Montenegro, 10-14 June, 2025, pp. 251-254, DOI: 10.1109/MECO66322.2025.11049201; also DOI: 10.48550/arXiv.2505.02060

[2] VideoFace2.0 demonstrations: https://github.com/brkljac/VideoFace2.0

Appendix A

Table A1 – List of files that are accompanying part of this report:

`VideoData_list.txt` Internet URLs of 2400 videos in VideoBase 1.0 collection.
`VideoData_list_IDs.txt` Unique identifiers of 2400 videos in VideoBase 1.0 collection.
`videoCapture.yml` Configuration file for VideoCapture software development environment.
`load_videoCapture_env.bat` Batch script for automated loading of VideoCapture environment (FFmpeg + yt-dlp libraries).
`video_infos_Dnevnik_RTV.txt` Detailed information about news broadcast videos from “Radio-televizija Vojvodine” (RTV).
`video_enc_Dnevnik_RTV.txt` Video and audio encoding characteristics of “Dnevnik_RTV” sample videos.
`video_ids_Dnevnik_RTV.txt` List of “Dnevnik_RTV” video identifiers.
`video_urls_Dnevnik_RTV.txt` List of “Dnevnik_RTV” video urls.
`video_infos_Dnevnik_N1.txt` Detailed information about news broadcast videos from “N1”.
`video_enc_Dnevnik_N1.txt` Video and audio encoding characteristics of “Dnevnik_N1” sample videos.
`video_ids_Dnevnik_N1.txt` List of “Dnevnik_N1” video identifiers.
`video_urls_Dnevnik_N1.txt` List of “Dnevnik_N1” video urls.
`video_infos_Dobro_Jutro_TANJUG.txt` Detailed information about discussion talk show videos from “TANJUG”.
`video_enc_Dobro_Jutro_TANJUG.txt` Video and audio encoding characteristics of “Dobro_Jutro_TANJUG” sample videos.

`video_ids_Dobro_Jutro_TANJUG.txt` List of “Dobro_Jutro_TANJUG” video identifiers.
`video_urls_Dobro_Jutro_TANJUG.txt` List of “Dobro_Jutro_TANJUG” video urls.
`video_infos_Uranak_K1.txt` Detailed information about news broadcast videos from “K1”.
`video_enc_Uranak_K1.txt` Video and audio encoding characteristics of “Uranak_K1” sample videos.
`video_ids_Uranak_K1.txt` List of “Uranak_K1” video identifiers.
`video_urls_Uranak_K1.txt` List of “Uranak_K1” video urls.
`videoCapture_stats.py` Video files duration statistics computation.

Appendix B

Table B1 – Overview of video and audio stream encoders for the internet video “ZAb0b1mhmrE”, video numbered 100 in the “video_ids_Dobro_Jutro_TANJUG.txt” list.

ID	EXT	RESOL.	FPS	CH	FILESIZE	TBR	PROT	VCODEC	VBR	ACODEC	ABR	MORE INFO
sb3	mhtml	48x27	0				mhtml	images				storyboard
sb2	mhtml	80x45	0				mhtml	images				storyboard
sb1	mhtml	160x90	0				mhtml	images				storyboard
sb0	mhtml	320x180	0				mhtml	images				storyboard
233	mp4	audio only					m3u8	audio only		unknown		Default, low
234	mp4	audio only					m3u8	audio only		unknown		Default, high
249-drc	webm	audio only		2	7.65MiB	50k	https	audio only		opus	50k	low, DRC, webm_dash
250-drc	webm	audio only		2	8.61MiB	57k	https	audio only		opus	57k	low, DRC, webm_dash
249	webm	audio only		2	7.65MiB	50k	https	audio only		opus	50k	low, webm_dash
250	webm	audio only		2	8.60MiB	57k	https	audio only		opus	57k	low, webm_dash
140-drc	m4a	audio only		2	19.69MiB	129k	https	audio only		mp4a.40.2	129k	medium, DRC, m4a_dash
251-drc	webm	audio only		2	15.68MiB	103k	https	audio only		opus	103k	medium, DRC, webm_dash
140	**m4a**	**audio only**		**2**	**19.69MiB**	**129k**	**https**	**audio only**		**mp4a.40.2**	**129k**	**medium, m4a_dash**
251	webm	audio only		2	15.67MiB	103k	https	audio only		opus	103k	medium, webm_dash
602	mp4	256x144	13		~13.01MiB	86k	m3u8	vp09.00.10.08	86k	video only		
269	mp4	256x144	25		~23.90MiB	157k	m3u8	avc1.4D400C	157k	video only		
160	mp4	256x144	25		8.90MiB	59k	https	avc1.4d400c	59k	video only		144p, mp4_dash
603	mp4	256x144	25		~25.43MiB	167k	m3u8	vp09.00.11.08	167k	video only		
278	webm	256x144	25		10.75MiB	71k	https	vp9	71k	video only,		144p, webm_dash
394	mp4	256x144	25		10.43MiB	69k	https	av01.0.00M.08	69k	video only		144p, mp4_dash
229	mp4	426x240	25		~42.56MiB	280k	m3u8	avc1.4D4015	280k	video only		
133	mp4	426x240	25		20.58MiB	135k	https	avc1.4d4015	135k	video only		240p, mp4_dash
604	mp4	426x240	25		~44.28MiB	291k	m3u8	vp09.00.20.08	291k	video only		
242	webm	426x240	25		18.47MiB	121k	https	vp9	121k	video only		240p, webm_dash
395	mp4	426x240	25		20.92MiB	138k	https	av01.0.00M.08	138k	video only		240p, mp4_dash
230	mp4	640x360	25		~76.56MiB	504k	m3u8	avc1.4D401E	504k	video only		
134	mp4	640x360	25		30.34MiB	200k	https	avc1.4d401e	200k	video only		360p, mp4_dash
18	mp4	640x360	25	2	82.13MiB	540k	https	avc1.42001E		mp4a.40.2	44k	360p

605	mp4	640x360	25		~93.79MiB	617k	m3u8	vp09.00.21.08	617k	video only	
243	webm	640x360	25		41.29MiB	272k	https	vp9	272k	video only	360p, webm_dash
396	mp4	640x360	25		36.19MiB	238k	https	av01.0.01M.08	238k	video only	360p, mp4_dash
231	mp4	854x480	25		~111.04MiB	731k	m3u8	avc1.4D401E	731k	video only	
135	mp4	854x480	25		60.11MiB	395k	https	avc1.4d401e	395k	video only	480p, mp4_dash
606	mp4	854x480	25		~142.64MiB	938k	m3u8	vp09.00.30.08	938k	video only	
244	webm	854x480	25		55.43MiB	365k	https	vp9	365k	video only,	480p, webm_dash
397	mp4	854x480	25		55.34MiB	364k	https	av01.0.04M.08	364k	video only	480p, mp4_dash
232	mp4	1280x720	25		~191.80MiB	1262k	m3u8	avc1.4D401F	1262k	video only	
136	mp4	1280x720	25		109.71MiB	722k	https	avc1.4d401f	722k	video only	720p, mp4_dash
609	mp4	1280x720	25		~228.65MiB	1504k	m3u8	vp09.00.31.08	1504k	video only	
247	webm	1280x720	25		102.14MiB	672k	https	vp9	672k	video only	720p, webm_dash
398	mp4	1280x720	25		91.73MiB	603k	https	av01.0.05M.08	603k	video only	720p, mp4_dash
270	mp4	1920x1080	25		~489.66MiB	3222k	m3u8	avc1.640028	3222k	video only	
137	**mp4**	**1920x1080**	**25**		**256.83MiB**	**1689k**	**https**	**avc1.640028**	**1689k**	**video only**	**1080p, mp4_dash**
614	**mp4**	**1920x1080**	**25**		**~392.62MiB**	**2583k**	**m3u8**	**vp09.00.40.08**	**2583k**	**video only**	
248	**webm**	**1920x1080**	**25**		**183.33MiB**	**1206k**	**https**	**vp9**	**1206k**	**video only**	**1080p, webm_dash**
399	**mp4**	**1920x1080**	**25**		**145.72MiB**	**958k**	**https**	**av01.0.08M.08**	**958k**	**video only**	**1080p, mp4_dash**
616	mp4	1920x1080	25		~845.07MiB	5560k	m3u8	vp09.00.40.08	5560k	video only	Premium