Adding Live captions

In this article we provide examples of calls and responses relative to Live Captions jobs in Reach integration flows.

Retrieving a Live captions job

Example of a response after requesting a Live caption job using entryVendorTask.getJobs - Kaltura VPaaS API Documentation:


```
{
  "objects":[
    {
      "id":"293434661",
      "partnerId":4428312,
      "vendorPartnerId":4728582,
      "createdAt":1667998823,
      "updatedAt":1667998823,
      "queueTime":1667998823,
      "entryId":"1_s8ms887c",
      "status":1,
      "reachProfileId":181162,
      "catalogItemId":12502,
      "price":0,
      "userId":"test@kaltura.com",
      "accessKey":"djJ8NDQzODM5Mnw6Jbim0ZsrnncsrmdDuMsd4iLBMCY3BS42rWx5qW_ZZJubaGXy-qkj2QuAjSpEOrH286GCnfbSF
      "version":0,
      "creationMode":1,
      "taskJobData":{
        "startDate":1668002400,
        "endDate":1668009600,
        "scheduledEventId":31111692,
        "entryDuration":7200000,
        "objectType":"KalturaScheduledVendorTaskData"
      },
      "expectedFinishTime":1668603623,
      "serviceType":2,
      "serviceFeature":8,
      "turnAroundTime":-1,
      "objectType":"KalturaEntryVendorTask"
    }
  ],
  "totalCount":1,
  "objectType":"KalturaEntryVendorTaskListResponse"
}
```


Delivering a Live captions job

Websocket response schema

Responses returned by the Streaming Speech Recognition services should have the following schema:

```
{
  "response": {
    "id": string (UUID),
    "type": "transcript" | "captions",
    "start": float,
    "end": float,
    "start_pts": float,
    "start_epoch": float,
    "is_final": boolean,
    "is_end_of_stream": boolean,
    "speakers": [
      {
        "id": string (UUID),
        "label": string | null
      }
    ],
    "alternatives": [
      {
        "transcript": string,
        "start": float,
        "end": float,
        "start_pts": float,
        "start_epoch": float,
        "items": [
          {
            "start": float,
            "end": float,
            "kind": "text" | "punct",
            "value": string,
            "speaker_id": string (UUID)
          }
        ]
      }
    ]
  }
}
```

Fields description

- `"response"` - The root element in the response JSON
  - `"id"` - A unique identifier of the response (UUID)
  - `"type"` - The response type. Can be either "transcript" or "captions" (See explanation in below note).
  - `"start"` - The start time of the utterance. Measured in seconds from the beginning of the media stream.
  - `"end"` - The (current) end time of the utterance. Measured in seconds from the beginning of the media stream.

- `"start_pts"` - The pts value corresponding to the `"start"` of the response, as received from the input media stream. Measured in seconds.
    - Note: if the input media stream doesn't provide pts values, this field will have the same value as `"start"`.
- `"start_epoch"` - The epoch timestamp at which the media corresponding to the `"start"` of the response was received.
- `"is_final"` - A boolean denoting whether the response is the final one for the utterance (See explanation in below note). For a "captions" response, this is always set to `"true"`, since captions are not incrementally updated (thus, each "captions" response is final).
- `"is_end_of_stream"` - A boolean denoting whether the response is the last one for the entire media stream
- `"speakers"` - A list of objects representing speakers in the media stream, as identified by the speech recognition service.
    - `"id"` - A unique identifier of the speaker (UUID)
    - `"label"` - A string representing the speaker. Only available in sessions with human transcribers in the loop. This field is set to `null` by default.
- `"alternatives"` - A list of alternative transcription hypotheses. At least one alternative is always returned.
    - `"transcript"` - A textual representation of the alternative in the current response.
    - `"start"` - Same as `["response"]["start"]`.
    - `"end"` - Same as `["response"]["end"]`.
    - `"start_pts"` - Same as `["response"]["start_pts"]`.
    - `"start_epoch"` - Same as `["response"]["start_epoch"]`.
    - `"items"` - A list containing textual items (words and punctuation marks) and their timings.
        - `"start"` - The start time of the item. Measured in seconds from the beginning of the media stream.

- `"end"` - The end time of the item. Measured in seconds from the beginning of the media stream.
- `"kind"` - The item kind. Can be either "text" or "punct" (a punctuation mark).
- `"value"` - The item textual value
- `"speaker_id"` - The unique identifier of the speaker that this item is associated with. Corresponds with an `"id"` of one of the speakers in the `"speakers"` field.

There are two types of responses - "transcript" and "captions":

1. **Transcript**: this type of response contains the recognized words since the beginning of the current utterance. Like in real human speech, the stream of words is segmented into utterances in automatic speech recognition. An utterance is recognized incrementally, processing more of the incoming audio at each step. Each utterance starts at a specific start-time and extends its end-time with each step, yielding the most updated result. Note that sequential updates for the same utterance will overlap, each response superseding the previous one - until a response signaling the end of the utterance is received (having `is_final == True`). The `alternatives` array might contain different hypotheses, ordered by confidence level.
2. **Captions**: this type of response contains the recognized words within a specific time window. In contrast to the incremental nature of "transcript"-type responses, the "captions"-type responses are non-overlapping and consecutive. Only one "captions"-type response covering a specific time-span in the audio will be returned (or none, if no words were uttered). The `is_final` field is always `True` because no updates will be output for the same time-span. The `alternatives` array will always have only one item for captions.

Responses on silent audio segments

It should be noted that "transcript" and "captions" responses behave differently when the audio being transcribed is silent:

- "transcript" responses are sent regardless of the audio content, in such a way that the entire audio duration is covered by "transcript" responses. In case of a silent audio segment, "transcript" responses will be sent with an empty word list, but with timestamps which mark the portion of the audio that was transcribed.
- "captions" responses are sent only when the speech recognition output contains words. In case of a silent audio segment, no "captions" responses will be sent, since a caption doesn't make sense

without any words. Therefore, "captions" responses will not necessarily cover the entire audio duration (i.e. there may be "gaps" between "captions" responses).

[template("cat-subscribe")]