dotsub™

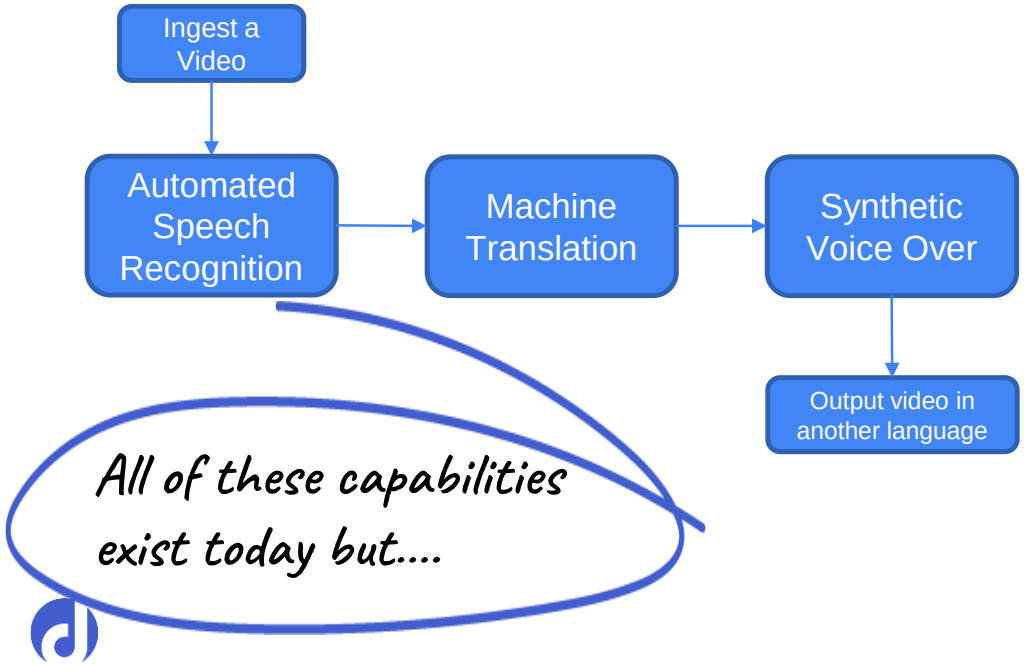
Machine Assistance in the Real World

A look at the real world of automation

Dave Bryant - Dotsub

September 2022

The ideal world



Ideal Video for automation

01. Single Speaker

02. Good Audio Quality

03. No background or ambient noise

04. Little or no jargon

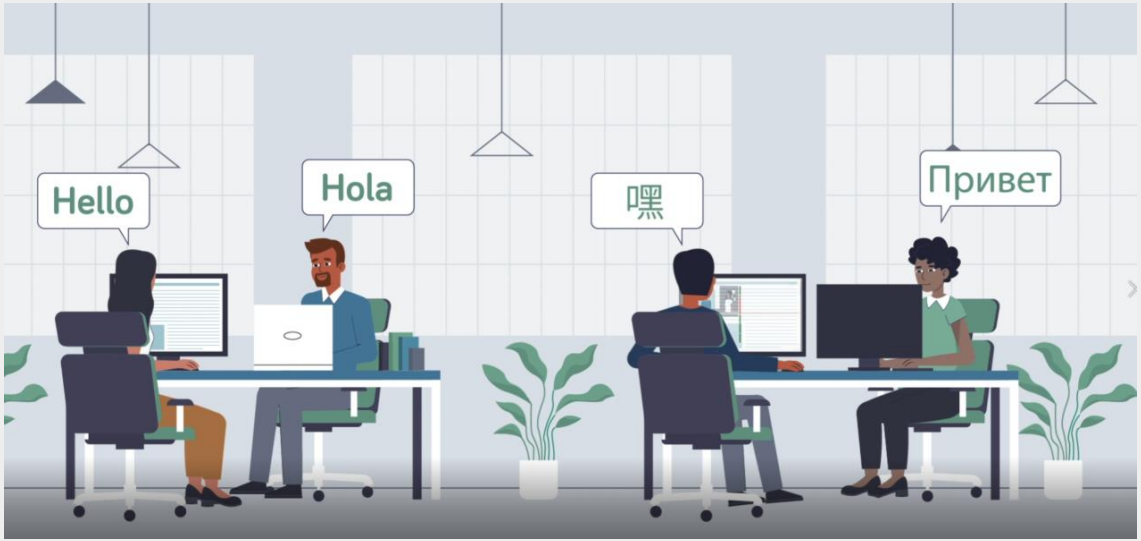
05. Relatively short sentences

06. Simple, clear Language

For this demonstration I will use a Dotsub explainer video. It is designed to be

1) Clear, concise and easily understood by all
2) Jargon Free
3) Excellent audio quality
4) Single Speaker with good diction

Play the 2-minute video



The foundation of this process are the English Captions

The captions need to be transcribed correctly.

- The only errors should be with proper nouns and names

They should be timed correctly

- No captions on the screen for too long or not long enough
- Should not extend over scene changes

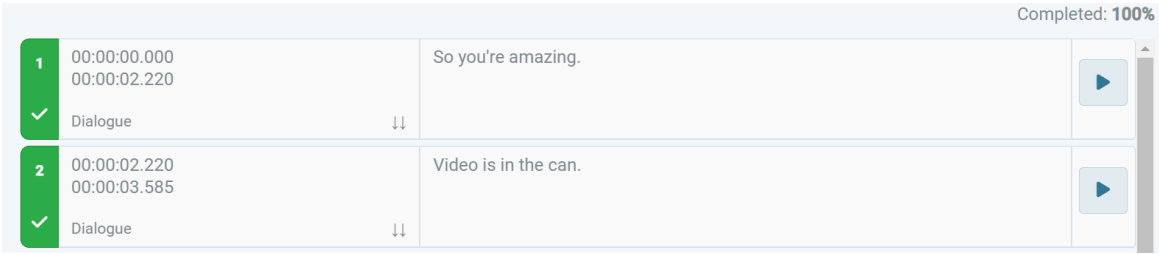
They should be well segmented

- The captions that are on the screen need to be logically grouped
- Should be comfortable to read

Let's run it through the ASR engine

The first line of the video's dialogue is "Your awesome video is in the can"

The ASR engine gives



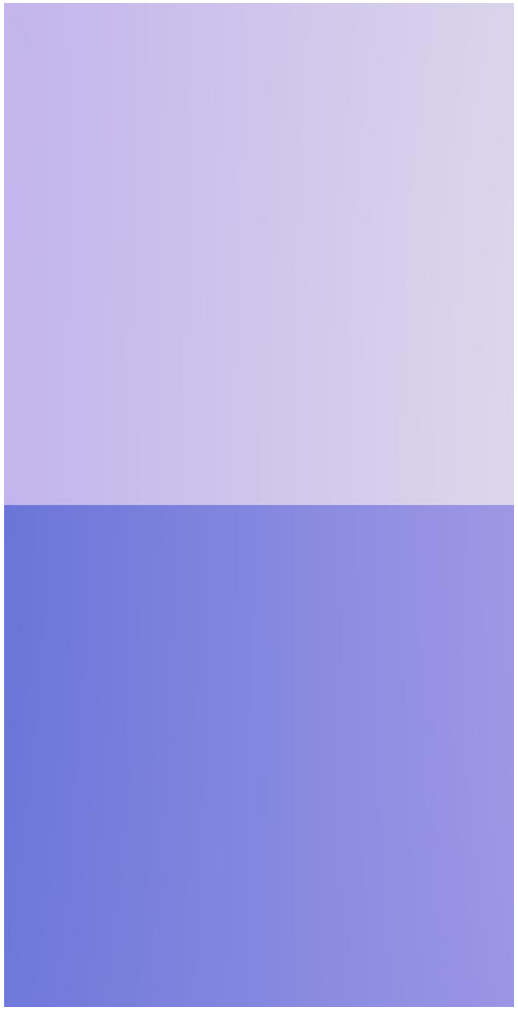
Not a great start.

Let's run it through the ASR engine (continued)

Other errors

Should be "Not so. Welcome to Any Video, Any Language from Dotsub."
ASR gave "Not so welcome to any video. Any language from dot sub."

Many examples of poor segmentation and therefore poor timing.



Comparing human captioner to ASR

```
1
- 00:00:00.000 --> 00:00:02.220 align:center
- So you're amazing.

2
- 00:00:02.220 --> 00:00:03.585 align:center
- Video is in the can.

3
- 00:00:03.585 --> 00:00:06.810 align:center
- It conveys the message you want to send perfectly.

4
- 00:00:06.810 --> 00:00:11.355 align:center
- And the dialogue is just right with loads of appeal for your native audience.

5
- 00:00:11.355 --> 00:00:15.315 align:center
- But what happens when your video gets seen and heard by non-Native viewers?

6
- 00:00:15.315 --> 00:00:18.615 align:center
- In fact, why limit yourself to one language

7
- 00:00:18.615 --> 00:00:22.470 align:center
- when you can reach a global audience in any language.

8
- 00:00:22.470 --> 00:00:28.080 align:center

- But surely that level of translation would take ages and cost a small fortune
```

```
1
+ 00:00:00.796 --> 00:00:03.837 align:center
+ So your amazing video
+ is in the can.

2
+ 00:00:03.837 --> 00:00:07.103 align:center
+ It conveys the message
+ you want to send perfectly,

3
+ 00:00:07.103 --> 00:00:09.727 align:center
+ and the dialogue is just right
+ with loads of appeal

4
+ 00:00:09.727 --> 00:00:11.688 align:center
+ for your native audience,

5
+ 00:00:11.688 --> 00:00:13.098 align:center
+ but what happens
+ when your video

6
+ 00:00:13.098 --> 00:00:15.683 align:center
+ gets seen and heard
+ by non-native viewers?

7
+ 00:00:15.683 --> 00:00:18.704 align:center
+ In fact, why limit yourself
+ to one language

8
+ 00:00:18.704 --> 00:00:20.542 align:center
+ when you can reach
+ a global audience
```

Most cues are very different

ASR Engines

We have the choice between 3 general purpose engines (as of August 2022)

All have their pros and cons.

We discourage the use of ASR without PE if translation is needed.

When using for translation the difference of speed and cost between human and ASR+PE

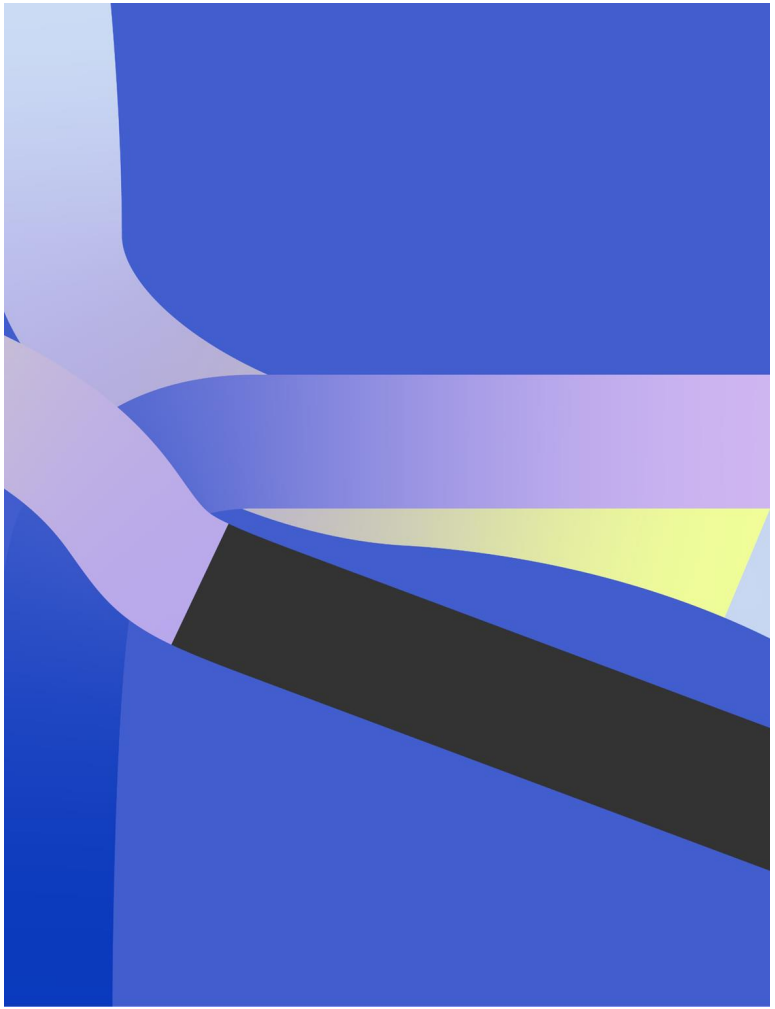
Machine Translation

MT for AVT is more difficult as a translation segment may be split across more than one cue

To maintain context you need to intelligently combine cues to make sure the correct concepts are translated

Once the translation is done then the timing and segmentation needs to be reapplied.

If used with excellent input (captions), then MT with light post editing works well



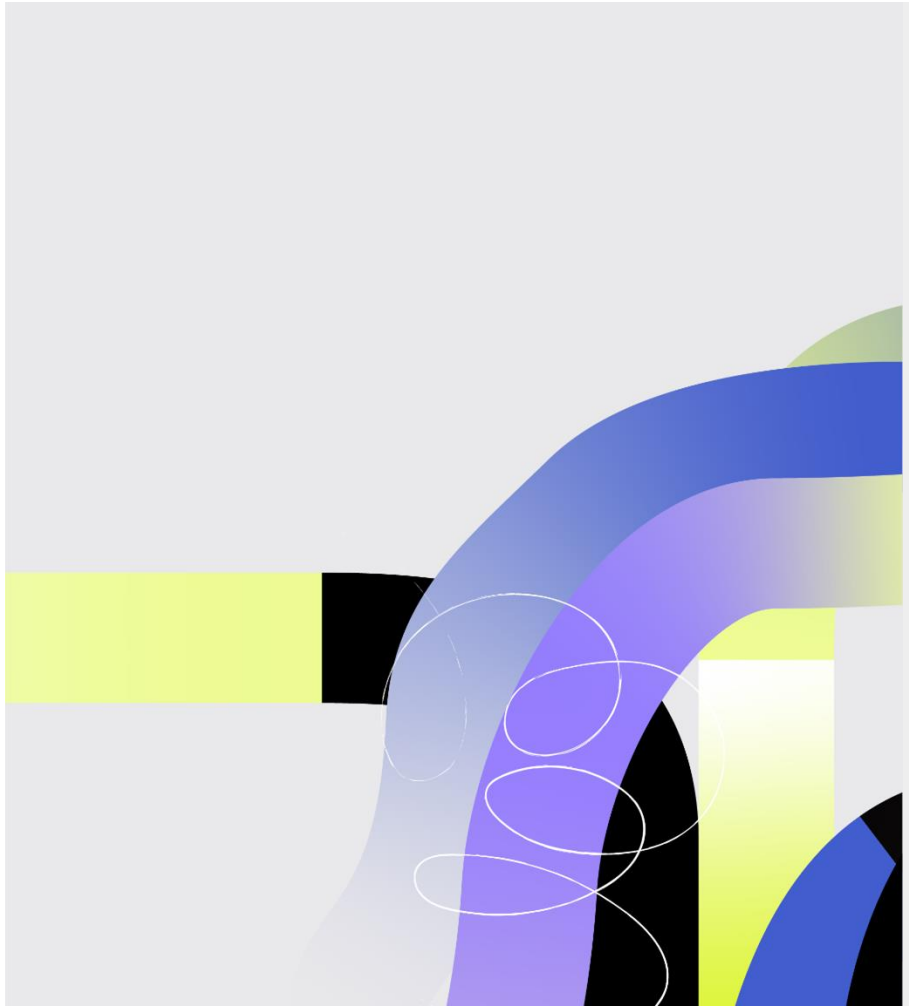
Synthetic Voice Overs – text to speech

This is the most exciting aspect of the whole scenario

Neural voices are generally very human like when used to voice videos that do not have a lot of emotional range. Good for explainer videos, how to videos, training videos and less
useful for dramatic entertainment videos.

Demonstrate a few voices to show quality

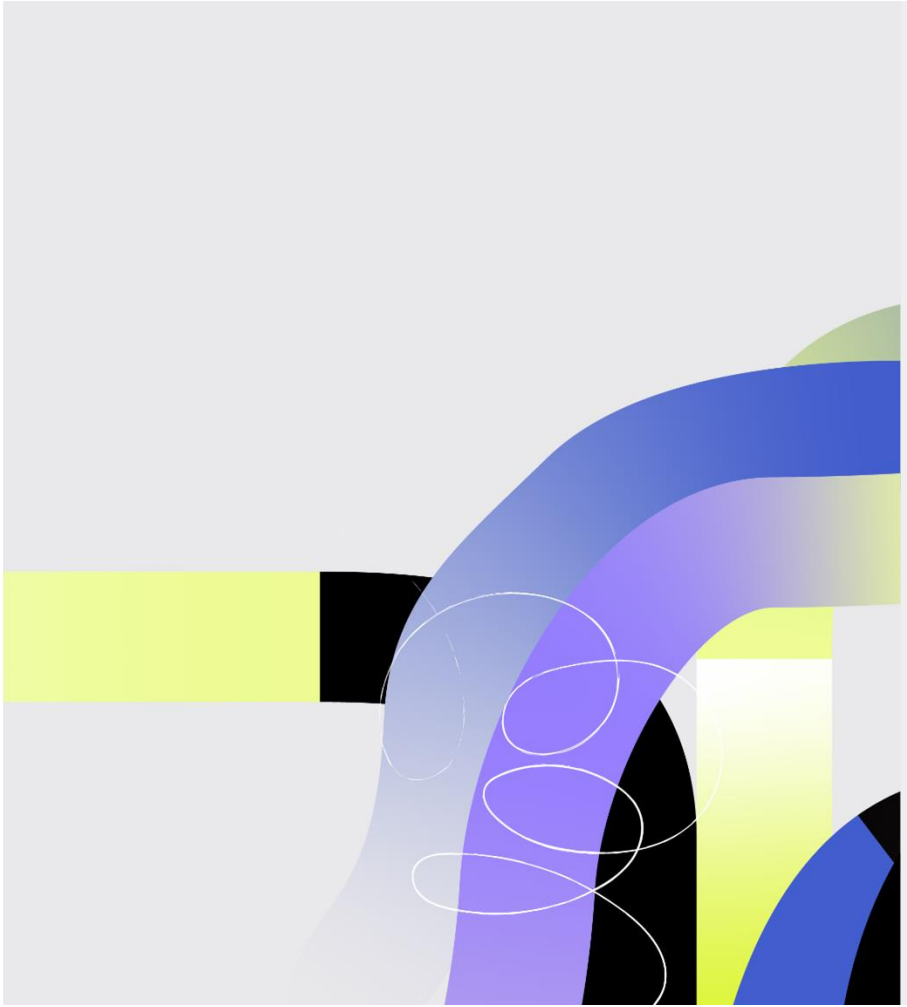
Currently, Microsoft Azure Cognitive Services provides 87 languages, each language having at least a male and female versión, more common languages have multiple dialects and speakers



Synthetic Voice Overs (continued)

Functionality includes

1) Fully automated workflow
2) Speaker ID and multiple voice support
    1) User can designate different voices to different speakers in the original video
3) No limit to the length of a SVO video
    1) Overcome limits of vendors
4) Videos synced with videos using the timing of the captions
    1) Long and short languages dealt with.
5) Editor within the platform that allows the prosody, emphasis and pronunciation of the SVO to be modified
6) Voiceover burnin
    1) The ability to demux the audio track so that the voice track is replaced while keeping the background audio (music or ambient)
7) Ability to create custom voices



Where we are today

Automation works but needs to be used cautiously
ASR often needs heavy postediting
MT only needs light postediting
Synthetic Voice Over is excellent in some situations

As of Q3 2022 – tomorrow, who knows?

We will provide examples of SVO's in multiple languages and dialects.

Thanks!

Dave Bryant

CEO, Dotsub

dave.bryant@dotsub.com

https://dotsub.com