THE BROADCAST BRIDGE PLAYOUT, CONTRIBUTION & DISTRIBUTION

Improving accessibility of content through Video Description

By Screen Subtitling Systems

Accessibility of media to partially sighted audiences may be enhanced by provision of additional information in the form of audio narrative track. An audio description of the video content is called "audio description" or "video description" in country-dependent nomenclature.

Traditionally, a trained 'describer' identifies appropriate points in the audio timeline where a description is needed (and can be placed) and produces a script. This process has much in common with captioning, but perhaps for historical reasons is almost always done separately. Typically, the 'describer' will also record the individual voiced segments, although sometimes the description is performed by a separate 'voice talent'.

Recent research indicates that users would be happy with computer-generated speech for Video Description if it leads to more provision. Where provision is being increasingly mandated, the economics may suggest that

Text To Speech is a viable alternative to using voice talents, particularly for non-premium channels.

The recorded audio segments are referenced from the script, and then this is typically used to produce a full-length audio track for mixing with the original audio. The description track may be 'pre-mixed' to create a separate audio track in advance, or live mixed with the original audio at play-out.

Sponsored by Screen Subtitling Systems

SCREEN SYSTEMS

WP WHITE PAPERS

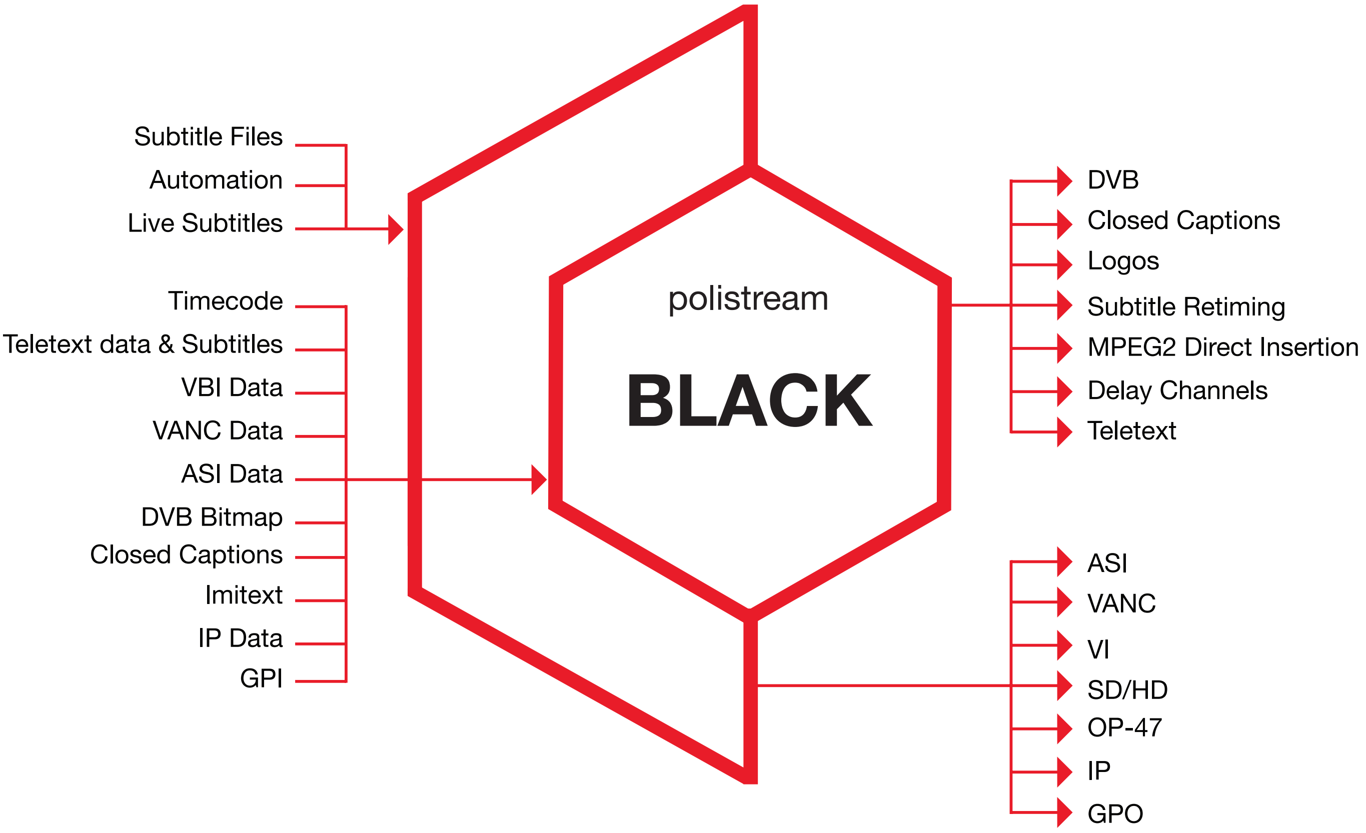

Figure 1. The Screen Polistream output driver drives a speech-to-text engine.

Live mixing typically uses a mono description track and a control track. This control track (sometimes termed a 'warble track' because of its sound) contains a low rate digital signal encoding pan and fade information that defines how the descriptive audio should be mixed with the original audio. This allows the balance between the description and the original audio to be controlled. Recent research indicates that users would be happy with computer-generated speech for Video Description if it leads to more provision. Where provision is being increasingly mandated, the economics may suggest that Text To Speech is a viable alternative to using voice talents, particularly for non-premium channels.

Spoken Subtitles

National broadcasters in some European countries currently provide spoken subtitles as a service addressing accessibility for the blind and partially sighted. Video description is also a requirement for these users, as Spoken Subtitles derived from subtitle files only replace the inaccessible foreign language narrative, but do not address the issue of missing visual cues. Unlike Video Description, Spoken Subtitles are traditionally provisioned using Text To Speech. This is because the textual data is already available in the form of subtitle files, and so adding machine reading of these is not operationally challenging. It is highly unusual for Spoken Subtitles to receive any special preparatory effort (for example, to match the voice with the gender of the speaker).

It should be understood that the 'quality' requirements of Spoken Subtitles may be different to Video Description. They can be provided automatically for all programs which have translation subtitles by default, which in some regions and channels is ALL programs. The original spoken audio (in a foreign language) is still audible, and so carries hint information such as the mood and gender of speaker. The need for the original audio to be heard in addition to the Spoken Subtitles may influence the mechanism of delivery (i.e. muting the original audio would detract from the quality of the viewing experience as the audible cues would then be removed).

Technical Implementation

Both Spoken Subtitles and Video Description have a common root in a timed script file. From this timed script file, audio is created. The main difference between the two practices is the ‘typical’ method of audio creation (using ‘voice talents’ for Video Description and using Text To Speech for Spoken Subtitles). Additionally there is a difference between the information supplied in the accessibility audio track and the need to retain connected information in the existing audio for Spoken Subtitles. However, technically, both Spoken Subtitles and Video Description may be provisioned using a Text To Speech engine.

Live insertion at program playout

Screen has developed an output driver for our Polistream subtitle and caption transmission system, which connects internally in the same way as all other output encoders (see figure 1). This specialist Polistream module receives ‘subtitle texts’ and renders them using SAPI 5 to drive a Text to Speech engine to produce an audio snippet. If the duration of the rendered audio is too long, it will be re-rendered with a faster spoken rate, up to a maximum configured speed increase. The audio snippet is then presented for output when the ‘subtitle text’ goes ‘on-air’. If an audio snippet is queued for output but is queued behind more than 4 seconds of existing audio data, then the previous audio snippet will be cut with a fade over 50ms so that the audio does not get progressively late in abnormal conditions.

The audio data is presented either duplicated into both channels of the stereo pair, or as mono with the right track filled with control data indicating the fade and pan values for the snippet. Those pan and fade values may be set to configured defaults, or drawn from metadata in the subtitle text (if present). The control track may be used to control an external mixer.

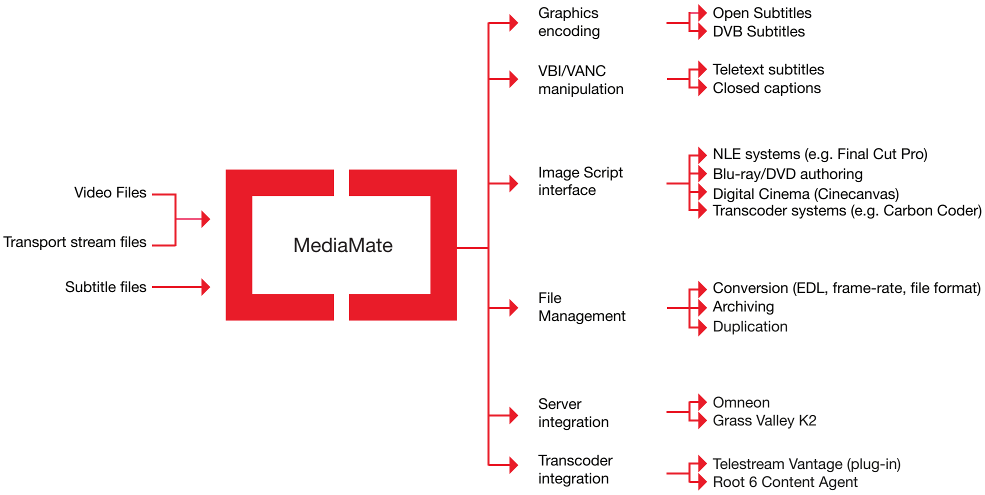

Figure 2. The Screen MediaMate spoken subtitles module.

The voices are specifically rendered at their native sample rate, and sample rate converted to 48 kHz for output to an AES-EBU output, which may be synchronized to a ‘world clock’ signal or to an incoming AES-EBU signal. Alternative delivery (e.g. Internet radio servers such as IceCast) can be facilitated by simple integration using virtual audio cables.

‘Live’ Video Description can also be rendered using the module, but in this case, the duration is unknown ahead of time, so there is no rate modification. The specialist module can also detect the presence of an audio filename as metadata in the subtitle, in which case the identified sound file is loaded instead of performing a Text To Speech operation. This allows for ‘voice-talent’ produced Video Description playout, a combination of Text To Speech and ‘voice-talent’ produced Video Description, or the playout of voiceovers or other short audio prompts.

Offline insertion processing framework

Screen has also developed a module for our MediaMate offline processing framework (see figure 2). The spoken subtitles module allows ‘subtitle text’ files to be rendered to 48 kHz stereo WAV files, with or without a control track. It behaves in a similar way to the Polistream implementation in terms of timing of audio and use of audio files. The generated audio file may then be simply attached to the media and played out alongside existing audio.