AUTOMATED VIDEO VOICE-OVER

THE HIDDEN PITFALLS OF AUTOMATED VIDEO VOICE-OVERS

INTRODUCTION:

Almost every localization company offers video voice-over services, with some even providing fully automated solutions. Before choosing the service that best suits your needs, it's crucial to be aware of potential pitfalls that may not be apparent until it's too late.

By reading this paper, you'll learn to ask the right questions to ensure you spend your video voice-over budget wisely and receive a service that aligns with your video communication goals.

KEYWORDS:

Voice-overs, Word Error Rate, AI Translation, Text-to-speech, artificial intelligence, artificial voices, neural voices

---

Consider this three

FULLY AUTOMATED ORIGINAL VOICE TRANSCRIPTION

Major players claim a Word Error Rate (WER) of about 5%. However, this can significantly increase in real-world scenarios due to factors like accents, background noise, overlapping speech (crosstalk), and specialized jargon, including your own brand name. Therefore, at a minimum, 5% of your content will be incorrectly transcribed, and likely more.

AI TRANSLATION

While AI translation works reasonably well for everyday conversations or social media posts, real-world scenarios involving context, idiomatic expressions, specialized jargon, and highly technical content can lead to higher error rates.

AI-BASED TEXT-TO-SPEECH (TTS)

Although AI TTS sounds impressive at first, users quickly encounter issues with numbers, dates, symbols, abbreviations, acronyms, and context-dependent pronunciations.

NATURAL LANGUAGE EXPANSION

Another important issue is the natural expansion of language when translating from English to another language. The translated speech often becomes longer. To match the length of the original audio, the solution typically involves speeding up the voice. This sounds unnatural and, in cases such as demos or tutorials, the video may play too fast for the audience to comprehend the content fully.

CONCLUSION

While AI-driven fully automated processes are cost-effective, their accuracy can be problematic. Without human intervention, a process with three automated steps (voice transcription, translation, text-to-speech), each generating a 5% error rate, results in an overall error rate of approximately 14.26%. This compounded error rate is significant and will likely affect the most crucial parts of your content.

Additionally, each language having different length and thus not matching the screen content can render your video hard to understand or unusable.