Nordic Vision & Speech Symposium (NoViS) 2025 · 24 November 2025

Shot-Aware Forced Alignment of Speech Transcripts in Edited Video

Jussi Pelkonen, Mei Takahashi, Tomasz Wierzba
VideoFinder Research — Helsinki · Kadrowa Labs Nordics Oy, 00180 Helsinki

Abstract

Speech recognisers produce word timestamps relative to the audio stream, but edited video introduces offsets that make those timestamps disagree with the frames: AAC encoder priming, MP4 edit lists, variable frame rate from phones and re-timed exports. We show that on a sample of 12,000 user uploads, 31% have an audio–video offset larger than one frame and 7% larger than 120 ms. We propose a shot-aware aligner that combines container-level timeline reconstruction with a constrained forced alignment that snaps word boundaries to nearby shot cuts when acoustic evidence is weak. Median alignment error drops from 92 ms to 23 ms, and moment retrieval on speech-grounded queries improves by 6.8 points R@1.

About this document

This is the public extended abstract. The full report, including appendices, ablations and evaluation scripts, is available on request from research@videofinder.link. Code and data referenced in this work are listed at https://helsinki.videofinder.link/open-source.html and /datasets.html.

Cite as

```
@techreport{pelkonen2025shotaware,
  title = {Shot-Aware Forced Alignment of Speech Transcripts in Edited Video},
  institution = {VideoFinder Research, Kadrowa Labs Nordics Oy},
  year = {2025}, number = {VFR-TR-2025-11}
}
```