Abstract

In film post-production, efficient methods for re-recording a dialogue or dubbing in a new language require a precisely time-aligned text, with individual letters time-coded to video frame resolution. Currently, this time alignment is performed by experts in a painstaking and slow process.

To automate this process, we used CRIM’s largevocabulary HMM speech recognizer as a phoneme segmenter and measured its accuracy on typical film extracts in French and English. Our results reveal several characteristics of film dialogues, in addition to noise, that affect segmentation accuracy, such as speaking style or reverberant recordings. Despite these difficulties, an HMM-based segmenter trained on clean speech can still provide more than 89% acceptable phoneme boundaries on typical film extracts.

We also propose a method which provides the correspondence between aligned phonemes and graphemes of the text. The method does not use explicit rules, but rather computes an optimal string alignment according to an edit-distance metric.

Together, HMM phoneme segmentation and phonemegrapheme correspondence meet the needs of film postproduction for a time-aligned text, and make it possible to automate a large part of the current post-synch process.