
Abstract
In this paper, we present the approach we used to produce a training database from a set of recorded newscasts for which we had inaccurate transcriptions. These transcribed segments correspond to a set of prepared anchor texts and journalist stories, not necessarily in chronological order of their actual presentation. No segmental time boundary information is provided. Our main concern is thus to establish time marks that delimit the audio segments of the corresponding texts. To resolve this problem, we have developed a time marking procedure using our speech recognition engine. We obtain a segmentation accuracy of 80%.