Annotation of Heterogeneous Multimedia Content Using Automatic Speech Recognition

Marijn Huijbregts, Roeland Ordelman, and Franciska de Jong

University of Twente, Dept. of Electrical Engineering, Mathematics and Computer Science, P.O. Box 217, 7500 AE, Enschede, The Netherlands
{m.a.h.huijbregts, ordelman, fdejong}@ewi.utwente.nl
http://hmi.ewi.utwente.nl

Abstract. This paper reports on the setup and evaluation of robust speech recognition system parts, geared towards transcript generation for heterogeneous, real-life media collections. The system is deployed for generating speech transcripts for the NIST/TRECVID-2007 test collection, part of a Dutch real-life archive of news-related genres. Performance figures for this type of content are compared to figures for broadcast news test data.

1 Introduction

The exploitation of linguistic content such as transcripts generated via automatic speech recognition (ASR) can boost the accessibility of multimedia archives enormously. This effect is of course limited to video data containing textual and/or spoken content but when available, the exploitation of linguistic content for the generation of a time-coded index can help to bridge the semantic gap between media features and search needs. This is confirmed by the results of TREC series of Workshops on Video Retrieval (TRECVID). The TRECVID test collections contain not just video, but also ASR-generated transcripts of segments containing speech. Systems that do not exploit these transcripts typically do not perform as well as the systems that do incorporate speech features in their models, or to video content with links to related textual documents, such as subtitles and generated transcripts.

ASR supports the conceptual querying of video content and the synchronization to any kind of textual resource that is accessible, including other full-text annotation for audiovisual material. The potential of ASR-based indexing has been demonstrated most successfully in the broadcast news domain. Spoken document retrieval in the American-English broadcast news (BN) domain was even declared ‘a solved problem’ based on the results of the TREC Spoken Document Retrieval (SDR) track in 1999. Partly because collecting data to train recognition models for the BN domain is relatively easy, word-error-rates (WER)
below 10% are no longer exceptional [9], and ASR transcripts for BN content approximate the quality of manual transcripts, at least for several languages.

In other domains than broadcast news and for many less favored languages, a similar recognition performance is usually harder to obtain due to (i) lack of domain-specific training data, and (ii) large variability in audio quality, speech characteristics and topics being addressed. However, as ASR performance of 50% WER is regarded as a lower bound for successful retrieval, speech-based indexing for harder data remains feasible as long as the ASR performance is not below 50% WER, and is actually a crucial enabling technology if no other means (metadata) are available to guide searching.

For 2007, the TRECVID organisers have decided to shift the focus from broadcast news video to video from a real-life archive of news-related genres such as news magazine, educational, and cultural programming. As in previous years, ASR transcripts of the data are provided as an optional information source for indexing. Apart from some English BN rushes (raw footage), the 2007 TRECVID collection consists of 400 hours of Dutch news magazine, science news, news reports, documentaries, educational programmes and archival video. The files were provided by the Netherlands Institute for Sound and Vision [2]. (In the remainder this collection will be referred to as Sound and Vision data.)

This paper reports on the setup and evaluation of the speech recognition system (further referred to as SHoUT system [3] that is deployed for generating the transcripts that via NIST will be made available to the TRECVID-2007 participants. The SHoUT system is particularly geared towards transcript generation for the kind of heterogeneous, real-life media collections exemplified by the Sound and Vision data that will feature in the TRECVID 2007 collection. In other words, it targets adequate retrieval performance, rather than plain robust ASR.

As can be expected for a diverse content set such as the Sound and Vision data, the audio and speech conditions vary enormously, ranging from read speech in a studio environment to spontaneous speech under degraded acoustic conditions. Furthermore, a large variety of topics are addresses and the material dates from a broad time period. Historical items as well as contemporary video fall within the range. (The former with poorly preserved audio; latter with varying audio characteristics, some even without 'intended' sound, just noise).

To reach recognition accuracy that is acceptable for retrieval given the difficult conditions, the different parts of our ASR system must be made as robust as possible so that it is able to cope with those problems (such as mismatches between training and testing conditions) that typically emerge when technology is transferred from the lab and applied in a real life context. This is in accordance with our overall research agenda: the development of robust ASR technology that can be ported to different topic domains with a minimum effort. Only technology that complies with this last requirement can be successfully deployed for the wide range of spoken word archives that call for annotation based on speech content such as cultural heritage data (historical archives and

---

2 Netherlands Institute for Sound and Vision: [http://www.beeldengeluid.nl/](http://www.beeldengeluid.nl/)

3 SHoUT is an acronym for SpecH recognition University of Twente.