Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 47 (2018) 214–239

www.elsevier.com/locate/csl

Automatic quality estimation for ASR system combination ☆

Shahab Jalalvand[a,b,*], Matteo Negri[a], Daniele Falavigna[a], Marco Matassoni[a], Marco Turchi[a]

[a] *FBK-Fondazione Bruno Kessler, Trento, Italy*
[b] *University of Trento, Italy*

Received 26 February 2016; received in revised form 11 April 2017; accepted 8 June 2017
Available online 15 June 2017

Abstract

Recognizer Output Voting Error Reduction (ROVER) has been widely used for system combination in automatic speech recognition (ASR). In order to select the most appropriate words to insert at each position in the output transcriptions, some ROVER extensions rely on critical information such as confidence scores and other ASR decoder features. This information, which is not always available, highly depends on the decoding process and sometimes tends to overestimate the real quality of the recognized words. In this paper we propose a novel variant of ROVER that takes advantage of ASR quality estimation (QE) for ranking the transcriptions at "segment level" instead of: *i)* relying on confidence scores, or *ii)* feeding ROVER with randomly ordered hypotheses. We first introduce an effective set of features to compensate for the absence of ASR decoder information. Then, we apply QE techniques to perform accurate hypothesis ranking at segment-level before starting the fusion process. The evaluation is carried out on two different tasks, in which we respectively combine hypotheses coming from independent ASR systems and multi-microphone recordings. In both tasks, it is assumed that the ASR decoder information is not available. The proposed approach significantly outperforms standard ROVER and it is competitive with two strong oracles that exploit prior knowledge about the real quality of the hypotheses to be combined. Compared to standard ROVER, the absolute WER improvements in the two evaluation scenarios range from 0.5% to 7.3%.

Keywords: Automatic speech recognition; Quality estimation; System combination

1. Introduction

The application of ASR systems in our daily life is steadily increasing. Voice search engines, voice question answering, broadcast news transcriptions, video/TV programs subtitling, meeting transcriptions and spoken. dialog systems are just some of the many applications involving ASR technology. In such applications, the quality of transcriptions and the availability of fast and accurate automatic evaluation methods represent two crucial interconnected needs. Automatic ASR evaluation, indeed, does not only represent a way to assess system performance

☆ This paper has been recommended for acceptance by Roger Moore.

* Corresponding author at: FBK-Fondazione Bruno Kessler, Trento, Italy.
E-mail address: jalalvand@fbk.eu (S. Jalalvand), negri@fbk.eu (M. Negri), falavi@fbk.eu (D. Falavigna), matasso@fbk.eu (M. Matassoni), turchi@fbk.eu (M. Turchi).

http://dx.doi.org/10.1016/j.csl.2017.06.003
0885-2308/

a posteriori, but also a way to improve it. For instance, automatic assessment methods can be used to select audio data for unsupervised training (Lamel et al., 2001; Falavigna et al., 2016) and active learning of acoustic models (Riccardi and Hakkani-Tur, 2005; Facco et al., 2006) or, as in the case of this paper, to combine multiple transcription hypotheses into a single and more accurate one.

In order to synthetically obtain more accurate transcriptions, ASR systems diversity and complementarity have been exploited in different ways (Audhkhasi et al., 2014). The combination of multiple hypotheses coming from independent sources usually leads to significant improvement compared to the output of each individual system. ROVER, the most popular ASR system combination approach, performs hypothesis fusion by first building a word confusion network (CN) from the *1*-best hypotheses of the ASR systems entering the combination and then by selecting the best word in each CN bin via majority voting (Fiscus, 1997). Word confidence scores, when available, can be exploited to inform the process by setting up a weighted majority voting scheme. Although several enhancements of this general strategy have been successfully proposed, ROVER-based hypothesis combination methods still suffer from some limitations that this paper aims to overcome.

The first one depends on how they are generally implemented: the hypothesis combination process considers the first input candidate as a "skeleton" to align the other hypotheses in a greedy manner. For this reason, depending on the order in which the hypotheses are considered when feeding the algorithm, the resulting combination can show large variations in quality. This raises the need of reliable techniques for ranking the input hypotheses before starting the fusion process.

The second limitation relates to the granularity of the input fed into ROVER: in the standard setting, transcription hypotheses correspond to entire audio recordings, whose duration can span up to hours. However, when manipulating long utterance transcriptions, the skeleton used to initialize the process can feature significant local variations in terms of quality. As a consequence, it may happen that the worst hypothesis for an entire audio recording (*i.e.* globally) is the best one for one or more passages (*i.e.* locally). This raises the need of strategies for operating at higher levels of granularity (*e.g.* segments spanning over few seconds) in order to take full advantage of local quality differences between the input hypotheses.

The third limitation concerns the applicability of ROVER-based hypothesis combination techniques: results' quality significantly increases when the availability of confidence scores makes it possible to set up weighted majority voting schemes. However, having access to the ASR systems' inner workings is a rigid constraint that limits the applicability of hypothesis fusion to scenarios in which the input transcriptions are produced by known ASR tools. Often, however, the hypotheses come from "black-box" systems, without additional scores.[1] This raises the need of methods that are independent from confidence information, but still capable to achieve good results with simple frequency-based voting.

Finally, it is worth noting that the confidence scores proposed by previous ASR literature (Evermann and Woodland, 2000a; Wessel et al., 2001), even when applicable, only indicate how confident the system is about its own output. This can be a biased perspective (influenced by individual decoder features), producing scores that are not comparable across different systems. External and system-independent measures of goodness would represent a more reliable alternative when comparable and objective ASR quality judgements are required.

To cope with these issues, in Negri et al. (2014) we proposed a reference-free and confidence-independent ASR quality estimation (QE) method, in which a supervised regression model is used to predict the word error rate (WER) of automatically transcribed audio recordings. Experimental results in different evaluation settings showed that our QE predictions can closely approximate the true WER scores calculated over reference transcripts. Building on these positive results, in Jalalvand et al. (2015b) we applied ASR QE to inform system combination with ROVER, outlining the framework that this paper aims to extend and refine. Our study focused on comparing the standard *system-level* ROVER with an alternative *segment-level* strategy that uses ASR QE to rank the input hypotheses before starting the fusion process. Our method was applied to combine the transcriptions of English TED talks produced by the eight participants in the IWSLT2013 evaluation campaign.[2] Its results outperformed the standard ROVER (based on averaging the results over a large number of random rankings of ASR outputs) and significantly

[1] For instance, this is the case of the steadily increasing volume of Youtube videos for which automatic captions are provided by black-box ASR technology. In 2012, more than 157 million videos were already accessible with auto-captions in 10 languages (source: http://goo.gl/5Wlkjl).

[2] The International Workshop on Spoken Language Translation (IWSLT – http://workshop2013.iwslt.org/) is a yearly workshop associated with an open evaluation campaign on spoken language translation.