Inter-annotator Agreement on Spontaneous Czech Language

Tomáš Valenta[1], Luboš Šmídl[2], Jan Švec[3]

1 Introduction

The goal of this article is to show that for some tasks in automatic speech recognition (ASR), especially for recognition of spontaneous speech, the gold-standard annotation differs substantially among human annotators. In this paper we focused on the evaluation of inter-annotator agreement (IAA) and ASR accuracy in the context of imperfect IAA. We evaluated it on a part of our Czech Switchboard-like spontaneous speech corpus. This part was annotated by three parallel transcriptions from three different annotators. The results give us additional insights for understanding of ASR accuracy.

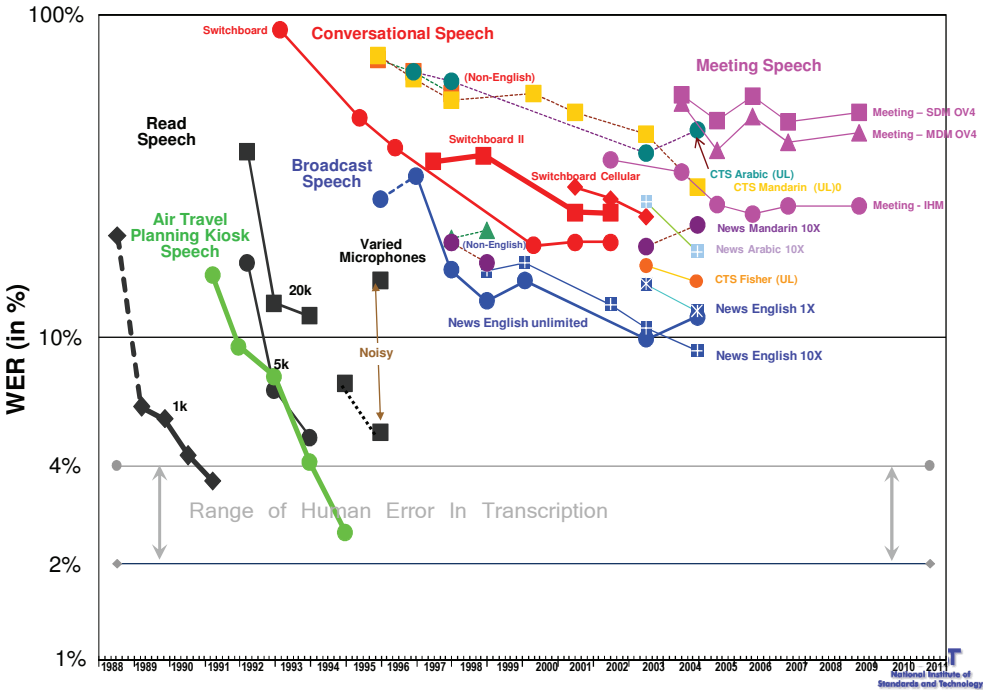

Figure 1: National Institute of Standards and Technology speech-to-text benchmark history, May 2009. Ajot and Fiscus (2009)

Automatic speech recognition accuracy differs significantly among various domains and tasks. On some tasks in some domains, the recognition accuracy almost attacks 100 %, whereas in others, it is about 60 % or less, as summarized Ajot and Fiscus (2009), see Fig. 1. Also, human transcription accuracy (i.e. IAA) above 90 % is almost unachievable in some tasks which sets upper bound for automatic speech recognition accuracy far below 100 %.

[1] student of doctoral study programme Applied Science and Informatics, field Cybernetics, e-mail: valentat@kky.zcu.cz

[2] assistant professor at the Department of Cybernetics, e-mail: smidl@kky.zcu.cz

[3] researcher at the Department of Cybernetics, e-mail: honzas@kky.zcu.cz

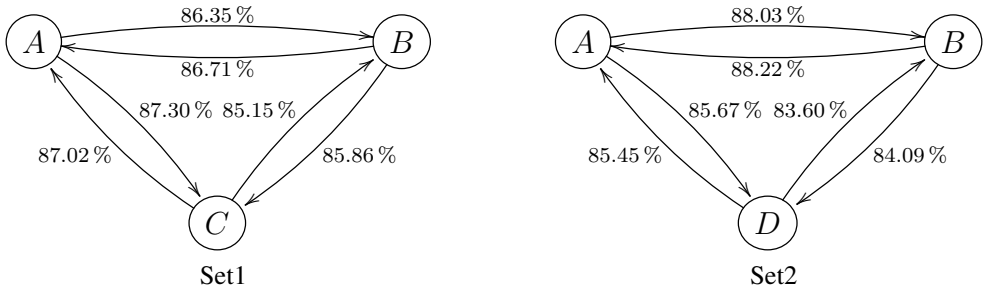

Figure 2: Inter-annotator agreement on evaluation sets. Annotator at the origin of the arrow was used as a reference and at the tip as a (recognition) hypothesis.

For this purpose Czech Switchboard-like corpus was chosen. It contains recordings of telephone communication of two people. The people usually know each other very well so they use lots of non-standard or local words and they speak colloquially. This reduces recognition performance significantly as well as the ability to recognize (and understand) by other people.

2 Inter-annotator agreement

Inter-annotator agreement was calculated using the same way as recognition accuracy, taking one annotation as a reference and the other as a recognition hypothesis and vice versa. First, the annotations were aligned so that the distance according to Levenshtein (1966) was minimal. Then the accuracy is calculated from the number of substitutions S, insertions I and deletions D of the alignment (N is the number of words in the reference):

$$Acc = \frac{N - S - I - D}{N}$$

Figure 2 shows IAA among three annotators on two evaluation datasets. Averaging the numbers in each subfigure and taking weighted average of them, we can estimate overall IAA on the corpus as 86 %.

3 Speech Recognizer Performance

Average speech recognition accuracy (taking the three annotations as a reference) is 49 % and 54 % on Set1 and Set2 respectively. ASR results are unquestionably worse than human transcription. Although it should be noted that ASR processes the audio in real time in a single pass. In contrast, the human annotator works about $8\times$ slower than real time, and also has the opportunity to play back the recording repeatedly.

Average (over the three reference annotations) "accuracy" of an annotator that heard the recordings just once is 70 % and 73 % on Set1 and Set2 respectively. His transcription rate was $5.5\times$ slower than real time. Recognizer accuracy with the one-pass reference annotation is 40 % and 46 % respectively.

References

Jérôme Ajot and Jonahtan Fiscus. Speech-To-Text (STT) and Speaker Attributed STT (SASTT) Results. *NIST Rich Transcription Evaluation Workshop*, 2009.

Vladimir Iosifovich Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, 1966.